REMARKS

Claims 1-24 are pending in this application. Claims 1-3 have been amended. Claims 4-21 and 23-24 have been withdrawn as the result of an earlier restriction requirement. Claim 22 has been cancelled. New claims 25-36 have been added. The specification has been amended to correct various typographical errors. The amendments and new claims do not add new matter. In view of the Office's earlier restriction requirement, Applicants retain the right to present claims 4-21 & 23-24 in a divisional application.

Amendments to the specification

Paragraph 0001 of the application is amended to clarify the status of this and related applications. This correction is a result of Applicant's cancellation of claim 22, which is claimed in the earlier application U.S. Ser. No. 10/436,376, filed May 12, 2003.

Also, various passages and two tables in the specification have been amended to correct typographical errors. All text and tables deleted in the current amendments correct errors that would be immediately obvious from reading the specification and Fig. 1. Specifically, the deleted text refers to molecular marker Satt228, which is located on Molecular Linkage Group (MLG) A2. The noted marker and MLG are incorrectly identified due to a typographical error. The specification and Fig. 1 clearly disclose the correct molecular markers that are associated with Rps8, (i.e. Satt 595, Satt516, Satt114, Satt334, Sat-317, Satt335, Satt510, Satt144, and Sat-197), all of which are on MLG F. See specification at Fig. 1, ¶0007, 0039, 0063, 0083 - 0085, Tables 6 & 7, and Examples 4 & 5.

The reason for the typographical errors is that the inventors' first attempt at localizing the novel Rps8 locus resulted in aberrant results that were described in two U.S. Provisional patent applications (*see* Burnham *et al.* (2003), U.S. Provisional Application No. 60/379,304, filed May

10, 2002; U.S. Provisional Application No. 60/427,637, filed November 19, 2002; and US Ser. No. 10/436,376, Filed: May 12, 2003). Upon discovering the errors reported in those applications, Applicants filed a new application and disclosed the correct methodology and molecular markers. The correct method and markers are the subject of the instant patent application; however, a few passages in the instant application were inadvertently duplicated from the earlier applications, thus resulting in inclusion of obviously incorrect information in the instant specification. Applicants hereby amend the specification to correct the inadvertent and erroneous information, specifically, references to Satt228 and MLG A2. Thus, the amendments do not add any new matter and are fully supported by the bulk of the specification, Fig. 1 and the working examples as stated above.

Claim rejection - 35 USC §112, second paragraph

The Office has rejected claim 3 as being indefinite. Claim 3 has been amended to more particularly point out and distinctly claim the subject matter which Applicants regard as the invention. Withdrawal of the rejection is respectfully requested.

Claim rejection - 35 USC §112, Enablement

Claims 1-3 have been rejected as failing to comply with the enablement requirement because, in the Office's view, "[t]he assignment of molecular markers to particular traits is unpredictable and population-specific." The Office has cited several references to support this statement.

The test of enablement is whether one reasonably skilled in the art could, without undue experimentation, make or use the invention from the disclosures in the patent coupled with information known in the art. (MPEP §2164.01). The factors to consider when determining whether a disclosure satisfies the enablement requirement and whether any necessary

experimentation is "undue" include: the amount of direction provided by the inventor; the existence of working examples; the level of one of ordinary skill; the state of the prior art; the breadth of the claims; the level of predictability in the art; the nature of the invention; and the quantity of experimentation needed to make or use the invention based on the content of the disclosure. (MPEP §2164.01(a)(citing *In re Wands*, 858 F.2d 731, 737, 8 USPQ2d 1400, 1404 (Fed. Cir. 1988)). It is improper to conclude that a disclosure is not enabling based on an analysis of only one of the above factors while ignoring one or more of the others. (MPEP *id.*) Applicants' analysis of the instant claims according to the *Wands* factors follows:

Like the inventors in *In re Wands*, the disclosure provides **considerable direction and guidance** on how to practice the invention as claimed in claims 1-3. The disclosure provides **working examples**. The **level of skill in the art** is high, requiring a practitioner to use molecular biology techniques, and the **state of the prior art** was such that all of the methods as well as the molecular markers needed to practice the invention were well known at the time the application was filed. (Information relating to the sequence of PCR primers to the 600+ SSR loci reported in Cregan et al. (1999) and a standard protocol for their amplification can be obtained on the USDA-ARS Soybean Genome Database, Soybase, at http://soybase.org/ (set-up in 1998, verified June 29, 2006.)

The claimed invention generally involves determining the presence or absence of *Phytophthora sojae* resistance in a soybean as indicated by the presence or absence of a newly-discovered resistance locus (Rps8), which maps to linkage group MLG F. (Specification at ¶0007.) According to the method, genomic DNA from a soybean is analyzed for the presence of the Rps8 locus. (*Id.*) The presence of the Rps8 locus is determined through the use of one or more molecular markers linked to Rps8. (*Id.*) This method is generally known as marker

assisted selection (MAS). (¶0028) Applicants provide the declaration of Dr. Anne Dorrance (Exhibit A) for an explanation of MAS.

Referring to the specification, Applicants first identified the presence of a new *P. sojae* resistance gene, Rps8, in the plant introduction PI 399073. (Specification at ¶0038.) Then, using routine and well known crossing and breeding methods, Applicants crossed the soybean plant carrying the Rps8 trait locus with a soybean plant having other specific desirable traits to produce a soybean line possessing the combination of desired phenotypes. (Specification at ¶ 0046, see generally ¶¶ 0043-0046.) Next, several crosses were created (HFX01-602, OX-98317, OX-99218, and OX-99128 disclosed in ¶¶ 0010-13 respectively) to produce progeny containing the Rps8 trait locus, as determined by their resistance to particular *P. sojae* pathotypes. (see specification at ¶0067 and Table 1, showing the resistance of OX-99218 to *P. sojae* pathotypes OH30 and OH4; ¶0071 and Table 3, showing the resistance of OX-99128 to *P. sojae* pathotypes OH30 and OH4; ¶0079 and Table 5, showing the resistance of the cross Williams x PI399073 to *P. sojae* pathotypes OH1 and OH25.). The seeds of one of the resultant germplasms, designated HFX01-602 was deposited with the ATCC in accordance with the Budapest Treaty. (Specification at ¶0074.)

Having determined the presence of Rps8 gene in a plant by phenotypic analysis, the genome of that plant was analyzed for the presence of markers associated with Rps8. (see ¶0081, explaining that the cross used for analyzing SSR marker association was Williams x PI399073; Example 2, generally explaining how the SSR markers linked to Rps8 were identified; and Example 4, the use of Joinmap and well known statistical analysis to determine which marker associations were statistically significant.) Applicants also confirmed their results using another cross. (Specification at ¶¶ 0089-93, Example 5.)

With respect to the MAS methodology, both the specification and publications in the art explain that the predictability of MAS increases if the number of markers is increased, and the markers bracket or flank the trait locus. (see specification at ¶ 0055; see also Demirbas et al (2001), p. 220, 2nd col., 2nd & 3rd ¶¶.) While claim 1 recites the use of at least two markers, there are several molecular markers disclosed in the specification, some of which flank the trait locus, and the number of markers known and available to those skilled in the art to be associated with the region of interest.

As Dr. Dorrance explains in her declaration (Declaration of Anne Dorrance, Page 2, Paragraphs 5 and 6), the inventors' own method, the extent of experience reported in the art with MAS, and the extent of information about molecular markers that was readily available to those skilled in the art as of the date the application was filed, was more than sufficient to enable a skilled artisan to successfully perform MAS for purposes of identifying soybean plants having the Rps8 locus.

Accordingly, Applicants maintain that the enablement requirement has been satisfied in view of the state of the art, together with the extent of disclosure in the specification regarding the Rps8 locus and various markers therefore.

As regards the **scope of the claims**, claim 1 recites soybean as the only genus in which the invention is practiced. (See specification at ¶0046, explaining that the invention may be applied generally to any plant variety of the genus *Glycine*, or soybean.) For this reason, the Office's reliance on Westman et al. (1997) ("Westman") is in error. Westman used molecular markers developed in the genus *Arabidopsis* to amplify marker loci in six *Brassica* crop species. (Westman, abstract, lines 5-8.) Westman, therefore, was evaluating whether markers developed

in one species could work "across taxa" in another species. This is not the method of the instant application, which is limited to molecular markers in only one species: soybean.

As regards **predictability of the art**, several factors negate the Office's conclusions about the unpredictability of MAS across different populations of soybean. First, the Office's reliance on Michelmore et al. (1991), van Ooijen et al. (1994) and Concibido et al. (1997) for the conclusion that these references "teach that it is unpredictable whether any particular PCR-derived or RFLP molecular marker developed with one population of soybeans may be successfully utilized with another population comprising the same species, or with interspecies hybrids" is misplaced. Michelmore, Concibido and Van Ooijen (as well as Lee et al. (1996)) only discuss non-SSR markers such as restriction fragment length polymorphism (RFLP) and random amplified polymorphic DNA primers (RAPDs). Not only are these references irrelevant as regards to claim 2 and 3, which recite SSR markers, but the teaching in these references cannot be extended to include "any particular PCR derived" molecular markers because these references specifically exclude SSR markers.

SSR markers are very different from RFLPs. This is because, as Cregan explains, only rarely have more than two alleles been identified as RFLP loci in soybean. (Cregan, p:1464, col. 2, second ¶.) Thus, because these two alleles generally have asymmetric frequencies, the likelihood that any two genotypes will be polymorphic at a particular RFLP locus is relatively low. (*Id.*) For this reason, using RFLPs alone, a polymorphic fragment mapped in one population may not be segregating in another. (*Id.*) A second drawback of using only RFLPs is the detection of multiple DNA fragments (i.e. multiple loci) with most probes. (*Id.*) The multiplicity of RFLP loci can make RFLP linkage maps ambiguous with respect to RFLP locus identity, and often

precludes the use of such loci for the evaluation of linkage group homology among different maps. (*Id* at p:1465, 1st ¶.)

In sharp contrast to RFLP markers, SSR markers are extremely useful because "the high levels of polymorphism, co-dominant inheritance, and the locus specificity of SSR markers in soybean" together with their "random distribution in the genome" make SSR markers "an excellent complement to RFLP markers for use in soybean molecular biology genetics and plant-breeding research." (Shoemaker et al. (1994)(Exhibit B), p. 241, 2nd ¶, lines 6-7 and 15-18. See also Song et al. (2004)(Exhibit C), p. 123, 1st Col., 1st ¶, lines 7-11: "Most SSRs are single-locus markers, and many SSR loci are multi-allelic. These characteristics make SSRs an ideal marker system not only for creating genetic maps, but also as an unambiguous means of defining linkage group homology across mapping populations.") Applicants refer to the declaration of Dr. Dorrance for an explanation of the fact that SSRs map consistently to the same genomic region across different soybean populations. (Declaration of Anne Dorrance, pages 2-3, Paragraphs 7 and 8)

With regard to the Lee and Concibido references cited by the Office, these references are simply inapplicable to the claimed invention because both Lee and Concibido used molecular markers to identify quantitative trait loci (QTLs) -- they were not in any way concerned with any single gene trait locus, such as Rps8. QTL's are not analogous to the Rps8 single gene locus of the present application because many genes affect QTLs (see, *e.g.* Lee at al, page 517, col. 1, second paragraph). In contrast, Rps8 is a single dominant gene, which segregates according to Mendelian genetic principles (Specification at ¶0046, 1st sentence; and ¶0078-79, disclosing a 3:1 resistant to susceptible ratio for the different families resulting from the Williams X PI399073, which is indicative that Rps8 segregated as a single dominant gene). This means that

1 in 4 progeny of a cross between a parent containing Rps8 and a parent that does not have Rps8 will have Rps8. Thus, there is a much higher probability (i.e. predictability) of successful MAS for a single gene locus as compared to a QTL. The context of the Lee and Concibido references are not analogous to that of the instant case, and one of ordinary skill in the art would find no motivation to use the methods described in these references for purposes of evaluating a single dominate gene locus.

Finally, similar to the analysis in *Wands*, the **nature of the technology** is such that it involves screening plants to determine which ones carry the desired Rps8 trait locus. Therefore, practitioners of this art are prepared to screen negative plants in order to find one that carries the desired trait. (See *In re Wands*, 8 USPQ2d 1400, 1406 (explaining that "the nature of monoclonal antibody technology is that it involves screening hybridomas to determine which ones secrete antibody with desired characteristics" and so "practitioners of this art are prepared to screen negative hybridomas in order to find one that makes the desired antibody" and that such screening did not constitute "undue experimentation.") Claim 1 has been amended for clarity to recite that detecting the presence of the molecular markers provides an indication that the trait locus Rps8 is present in the soybean. (See specification at Summary, ¶ 0007, first sentence.) It is well known that MAS is an "indirect" method of selection based on the probability that if a plant carries a marker associated with a particular trait locus, there is a probability that the plant carries the trait locus. With a higher probability of successful selection, based on a high level of statistical significance, there is a higher predictability that the methods will be useful.

"The presence of inoperative embodiments within the scope of a claim does not necessarily render a claim nonenabled." (MPEP §2164.08(b).) Claims reading on significant numbers of inoperative embodiments would render claims nonenabled only when the

specification does not clearly identify the operative embodiments and undue experimentation is involved in determining those that are operative. (*Id.*) The instant application clearly identifies the operative embodiments and, given the amount of guidance in the specification, the amount of experimentation required in determining the operative embodiments is not undue. This is because those skilled in the art tailor the molecular markers to the particular cross (or population) that they are interested in, thereby increasing the predictability of successful marker assisted selection. Several maps of related species, showing how regions can be reorganized, are known in the art and are readily available, for example on the soybase database referenced above, so that skilled artisans would be able to tailor markers to the population of interest. The test for undue experimentation is "not merely quantitative," and "a considerable amount of experimentation is permissible, if it is merely routine, or if the specification in question provides a reasonable amount of guidance with respect to the direction in which the experimentation should proceed." (*Ex parte Jackson*, 217 USPQ 804, 807 (1982)).

In view of the *Wands* factor analysis above, Applicants submit that the instant specification provides the requisite "reasonable amount of guidance" by specifically disclosing the plant phenotypes that have the desired and novel Rps8-associated *P. sojae* resistance trait, and by mapping the Rps8 locus to MLG F, to enable a skilled artisan to determine whether a soybean plant has Rps8 trait locus. Moreover, by providing SSR molecular markers that are associated with the Rps8 locus, two of which flank the locus, the predictability of the MAS is sufficiently increased to enable a person skilled in the art to make and/or use the invention as claimed. For the foregoing reasons, withdrawal of the enablement rejection is respectfully requested.

Claim rejection - 35 USC §102, Anticipation

Claims 1-3 have been rejected as being anticipated by Demirbas et al. (2001) ("Demirbas").

Amended claim 1 recites a method for determining the presence of trait locus Rps8 in a soybean by analyzing the genomic DNA of the soybean for the presence of at least two molecular markers associated with trait locus Rps8. Although Demirbas discloses Satt114, which Applicants discovered to be associated with Rps8, Demirbas, which is concerned with Rps3, does not disclose any other markers associated with Rps8. Since claim 1 recites at least two molecular markers associated with Rps8, Demirbas does not anticipate claim 1.

Claims 2 and 3 depend from claim 1 and so are novel for at least the same reasons as claim 1. Withdrawal of the rejection is respectfully requested.

Claim rejection - 35 USC §103, Obviousness

Claim 3 is rejected as obvious in view of Demirbas and Cregan et al. (1999). Applicants respectfully maintain that the Office has failed to make a *prima facie* case of obviousness.

According to the Office, a skilled artisan would have been motivated to combine the Satt114 marker, disclosed in Demirbas to be associated with Rps3, and Satt516 marker, disclosed in Cregan to be located on linkage group F, to arrive at claim 3. This, however, is a clear example of "improper hindsight" reconstruction of the invention. (MPEP §2142.) According to established law, although every element of a claimed invention may often be found in the prior art, identification in the prior art of each individual part claimed is insufficient to defeat patentability of the whole claimed invention. (*In re Kotzab*, 217 F.3d 1365, 1370 (Fed. Cir. 2000)) To establish obviousness based on a combination of the elements disclosed in the prior art, there must be some suggestion or motivation, either in the references themselves or in the

knowledge generally available to one of ordinary skill in the art, to modify the reference or to combine reference teachings. (MPEP §2143.01 I.)

There is absolutely no teaching, motivation, or suggestion either in Cregan or in Demirbas, to combine Satt114 and Satt516 to find a new *P. sojae* resistance locus on linkage group F. This is because first, a skilled artisan would have had to find a plant that possessed the required phenotype (resistance to *P. sojae* pathotypes vir1a, 1b, 1c, 1d, 1k, 2, 3a, 3b, 3c, 4, 5, 6 and 7) before finding the novel gene which conferred the particular resistance to *P. sojae*. Such a plant was not disclosed in either Demirbas or Cregan, and was not within the knowledge of one of ordinary skill in the art because it was not known prior to the Applicants' disclosure thereof. All soybeans would probably have a particular marker, such as Satt114, in their DNA; it is the association of that marker with the new phenotype from the source plant that is novel and non-obvious.

Additionally, Demirbas established that other *P. sojae* resistance loci Rps1 - Rps6 are located on linkage groups N, J, F, and G. Thus, absent considerable experimentation, a skilled artisan in possession of the disclosure by Demirbas would not have known which linkage group a new Rps locus would map to, and would not have been motivated to choose a molecular marker associated with linkage group F to combine with any other molecular marker.

Lastly, although Demirbas et al. disclosed Satt114 to be "moderately linked" to Rps3, the authors concluded that "Neither Satt114 nor Satt374 displayed any significant linkage to Rps3" and excluded Satt114 as a useful marker for efficient marker assisted selection for Rps3 (see Discussion, page 1226, 1st col., 2nd ¶, lines 2-7.) Thus, even if the location of the new Rps8 gene locus was found to be related to the location of Rps3, such a finding was not reported by Demirbas or in any other art as of the filing date of the instant case, and importantly, Demirbas

teaches away from using Satt114 for marker assisted selection of Rps3 and provides no teaching to use that marker for any other trait, much less Rps8.

Upon reading Demirbas and Cregan, one of ordinary skill would find no motivation to combine the teachings for the reasons given above. And even if the teachings of these references were combined, the combination would not provide the methods of the instant claims because, as noted above, the discovery of the novel Rps8 trait locus and the associated *P. sojae* resistance had not yet been made. For these reasons, Applicants maintain that claim 3 is not obvious. Withdrawal of the rejection is respectfully requested.

New claims 25 to 36 have been added. Support for claim 25 is found in the specification at ¶ 0066, ¶0070, and Example 1. Claim 26 is supported by the specification at ¶0038, and ¶¶0045-46. Support for claims 27 to 34 is found in the specification at ¶0007 ("The presence of the Rps8 gene is determined through the use of one or more molecular markers linked to Rps8"); ¶0028; ¶¶ 0050-0055; and ¶¶0056-0061. Support for claims 35 and 36 is found in the specification at ¶0067, ¶0071, ¶0079 and Tables 1, 3 & 5. The new claims do not add new matter.

It is respectfully submitted that the application is now in condition for allowance. Applicants respectfully request that a timely Notice of Allowance be issued in this case.

Respectfully submitted,

Date: September 5, 2006

By: /diane h. dobrea/
Diane H. Dobrea
Reg. No. 48,578
(614) 621-7788

EXHIBIT

A

Customer Number

24024

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of:	)	Group Art Unit: 1638
St. Martin *et al.*	)	Confirmation No.: 3349
Application No.: 10/778,018	)	Examiner: Keith O'Neal Robinson
Filed: February 12, 2004	)	Attorney Docket No.: 22727/04212
For: Identification of Soybeans Having Resistance to *Phytophthora Sojae*	)	

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313

Declaration of Dr. ANNE DORRANCE under 37 C.F.R § 1.132

I, Anne Dorrance, an inventor in the above-identified application, declare as follows:

1. I received a Ph.D. from Virginia Polytechnic Institute and State University and postdoctoral training at Washington State University. Since then, I have been employed as faculty in the Department of Plant Pathology at The Ohio State University. I have published 24 (3 more are in press) full-length publications in peer-reviewed international scientific journals.

2. Currently, I am an associate Professor of Plant Pathology at The Ohio State University.

3. I am a co-inventor of the above application, and have directed research relating to soybean plants for nine years. I have published 12 (3 more are in press) peer-reviewed papers on resistance to *Phytophthora sojae* in scientific journals.

4. I would like to comment on the process of marker assisted selection or MAS to clarify (a) our experiments, which form the basis of the patent application; and (b) how other scientists can successfully perform MAS based on the information in our application.

5. The steps that are used in MAS may be summarized as follows: (i) identifying the locus/gene of interest, i.e. Rps8, which confers the desired trait, i.e. a new, Rps-8 derived resistance to *P. sojae* pathotypes that will normally kill plants, including plants that have one or more of the previously identified Rps genes (Rps1a, Rps1b, Rps1c, Rps1d, Rps1k, Rps2, Rps3a, Rps3b, Rps3c, Rps4, Rps5, Rps6 or Rps7); (ii) cross breeding plants that have the Rps8 locus with plants that do not have the Rps8 locus to develop progeny segregating for the trait; (iii) identifying molecular markers that are genetically linked to the Rps8 locus/gene and mapping the Rps8 locus. MAS can then be performed on progeny developed from any cross with a parent that has the Rps8 gene to determine which progeny carry the Rps8 gene. The parent with the Rps8 gene is chosen based on its phenotypic characteristics. MAS is performed using molecular markers from the region that was identified in step (iii).

6. As explained in the specification, we carried out steps (i), (ii) and (iii) and mapped the Rps8 locus to a particular region on major linkage group (MLG) F. We also provide nine SSR molecular markers, two of which flank the Rps8 locus, which are genetically linked to the Rps8. SSR markers consistently map to the same region of the *Glycine* (soybean) genome. This means that in all soybean populations tested, SSR markers map to a single locus in the genome with a map order that is essentially identical in all populations. (Shoemaker et al. 2004 (Exhibit A), p.243, 2nd ¶, lines 13-16.) This has been demonstrated across a number of soybean populations and is the basis of the "consensus" map of the soybean genome (Cregan et al., 1999). In other words, as a result of the SSR cross-population consistency of mapping, Cregan, in 1999, was able to align the 20+ linkage groups derived from each of three soybean populations into a consensus set of homologous groups to correspond to the 20 pairs of soybean chromosome. Thus, the consensus map of the soybean genome was compiled

combining SSR data from several *Glycine* populations. All of this information (the SSR markers, the sequence of the PCR primers to the 600+ SSR loci reported in Cregan, a standard amplification protocol, and the consensus map) is available to the soybean community at http://soybase.org/.

7. Given the consistency of SSR data across soybean populations together with the information provided in the instant application, scientists in the field would be able to choose appropriate molecular markers for any cross that displays the Rps8-associated resistance phenotype. Moreover, any molecular marker that maps to the region on MLG F identified in our application as encompassing Satt516 to Satt114, and that is polymorphic for the parents in that region, can be used for successful MAS.

8. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

8-25-06
Date

Anne Dorrance
Dr. Anne Dorrance

EXHIBIT

B

6 Soybean Genomics

RANDY C. SHOEMAKER

Iowa State University
Ames, Iowa

PERRY B. CREGAN

Beltsville Agriculture Research Center
Beltsville, Maryland

LILA O. VODKIN

University of Illinois
Urbana, Illinois

6–1 THE SOYBEAN GENOME

Soybean [*Glycine max* (L.) Merr.] has emerged as a model crop system because of its densely saturated genetic map (Cregan et al., 1999), a well-developed genetic transformation system (Clemente et al., 2000; Xing et al., 2000; Zhang et al., 1999), and the growing number of genetic tools applicable to this biological system (reviewed in Shoemaker, 1999). It is also the number one oilseed crop in the world and a multibillion-dollar crop of the USA (Riley, 1999; SoyStats, 1997).

The soybean genome is of average size compared to that of many other plants. It is comprised of about 1.1 Mbp/C (Arumuganathan and Earle, 1991). This makes it about seven and one-half times larger than the genome of Arabidopsis and two and one-half times larger than rice (*Oryza sativa* L.). Still the soybean genome is less than half the size of the corn (*Zea mays* L.) genome and more than 14 times smaller than the genome of bread wheat (*Triticum aestivum* L.) (Arumuganathan and Earle, 1991). Approximately 40 to 60% of the soybean-genome sequence can be defined as repetitive (Gurley et al., 1979; Goldberg, 1978). One family of repetitive sequence, STR120, is comprised of an approximate 120 bp monomer (Morgante et al., 1997). This family is estimated to consist of 5000 to 10 000 copies. Some repetitive sequence families may be species-specific (Morgante et al., 1997).

Bacterial artificial chromosome (BAC) libraries (Marek and Shoemaker, 1997; Danesh et al., 1998; Tomkins et al., 1999; Salimath and Bhattacharyya, 1999; D. Lightfoot, personal communication, 2002) also have been produced, which together cover the soybean genome many times over. Detailed physical contigs have already been developed and reported using some of these libraries (Marek and Shoemaker, 1997). These libraries have been made from different genotypes

 Soybeans: Improvement, Production, and Uses, 3rd ed, Agronomy Monograph no. 16.

and with a variety of enzymes and most have been made available to the public. Yeast Artificial Chromosomes have also been created for the purpose of chromosome walking and in situ hybridization (Zhu et al., 1996).

The degree to which soybean chromosomes constrict during meiosis has made it difficult to conduct cytogenetic analyses. However, a complete karyotype has now been reported (Singh and Hymowitz, 1988) based upon pachytene analysis. Analysis of pachytene chromosomes has shown that more than 35% of the genome is made up of heterochromatin, with the short arm of six of the 20 bivalents being completely heterochromatic (Singh and Hymowitz, 1988).

Primary trisomics (genomic complements containing one additional chromosome; $2n = 41$) are useful for quickly locating genes onto a specific chromosome and for associating linkage groups with specific chromosomes. By using aneuploid lines that Dr. R. Palmer (USDA-ARS) supplied, and some generated at the Soybean Cytogenetics Lab at Urbana-Champaign, IL a complete set of 20 trisomics, in which each chromosome exists in an extra copy, is now completed (Xu et al., 1998). This work will undoubtedly be useful for integrating classical and molecular genetics, much as similar cytogenetic collections have been important for maize, rice, barley (*Hordeum vulgare* L.), and tomato (*Lycopersicum esculentum* Mill.).

Soybean has a diploid chromosome number of $2n = 40$. However, most genera in the *Phaseolae* have a genome complement of $2n = 22$ (Hymowitz et al., 1998). This led Lackey (1980) to suggest that *Glycine* was probably derived from a diploid ancestor ($n = 11$) which underwent aneuploid loss to $n = 10$ and subsequent polyploidization to yield the present $2n = 40$.

Some type of polyploidization event has very likely occurred in the soybean's distant past. However, in spite of being a polyploid, the genome, for the most part, acts like a diploid. The 'diploidization' of polyploids is a well-known process and is caused by additions, deletions, mutations, and rearrangements that rapidly inhibit nonhomologous pairing of linkage groups (Ohno, 1970). Examples of divergence of duplicated genes have been reported for soybean receptor-like protein kinases (Yamamoto and Knap, 2001) and CLV1-like genes (Yamamoto et al., 2000). However, this may be a relatively slow process and there remain many exceptions reminding us that soybean is a tetraploid.

Many examples of duplicate factor genes (two independent genes controlling the same trait) can be found in the soybean germplasm collection (Palmer and Kilen, 1987). An analysis of the average number of fragments detected by hybridization to restriction-digested soybean-genomic DNA by each of 280 randomly chosen *Pst*I genomic clones emphasized the abundant duplications found in the genome (Shoemaker et al., 1996). More than 90% of the probes detected more than two fragments and nearly 60% detected three or more fragments. This suggests that <10% of the genome may be single copy sequence and that large amounts of the genome may have undergone genome duplication in addition to the presumed tetraploidization event.

Another analysis of hypomethylated sequences using methylation sensitive restriction enzymes indicated that slightly more than 15% of the hypomethylated genome remains as single-copy DNA (Zhu et al., 1994). The remainder of the hypomethylated genome appeared to be duplicated or middle-repetitive sequence. No evidence of silencing of the duplicated regions was observed by methylation.

Hybridization-based mapping has resolved many duplicated regions of the genome (Cho et al., 1989). These homoeologous regions reflect segmental and whole-genome duplication events and can provide much information about the evolution of the genome (Fig. 6–1). Sequences are often duplicated in the soybean genome in a manner not easily explained by a tetraploidization event. For example, most linkage groups contain markers that can also be found on other linkage groups but examples of this can sometimes be extreme. For any given linkage group, duplicate markers may be present on more than one other linkage group. The average linkage group contains markers that can be found on eight other linkage groups (Shoemaker et al., 1996).

Mapping of duplicated genes controlling pubescence morphology (appressed and non-appressed) provided interesting insight into the evolution of the genome. Lee et al. (1999) mapped *Pa*1 and *Pa*2 to LG-B1/S and LG-F, respectively. It was expected that these genes would map to homoeologous segments of the linkage groups. However, other than the genes, no markers appeared in common between these regions. Unexpectedly, the gene regions were implicated as paralogs through an intermediate linkage group, LG-H, which connected to LG-B1/S and LG-F regions through multiple markers. This suggested that regions of LG-B1/S and LG-F (as well as LG-H) were evolutionarily related and that perhaps a third pubescence gene existed and remained undetected.

Except for many disease-resistance genes, most agronomically important traits are controlled by several to many genes acting in concert. The genetic locations of the quantitative gene(s) are known as quantitative trait loci (QTL). Because of the genetic by environment interactions on most quantitative traits, breeding for them requires replicated field trials conducted over 2 or more years in a variety of locations. This is obviously time consuming and expensive. The ability to select for an

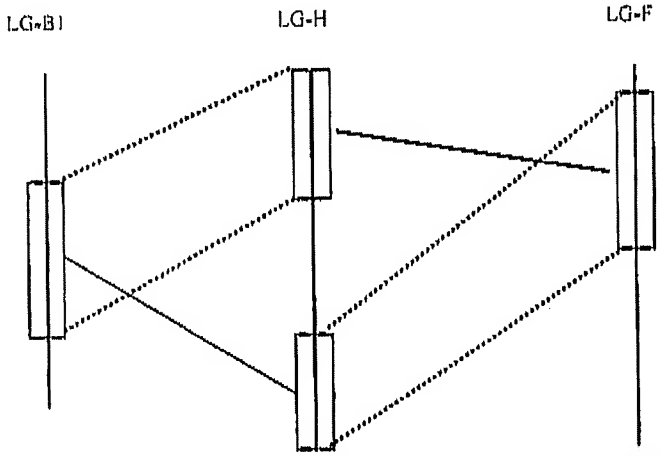

Fig. 6–1. Examples of homeologous regions in soybean detectable with hybridization-based mapping techniques. In this example LG-H has homeologs on both LG-B1 and LG-F, while segments of LG-B1 and LG-F are connected through a single marker with homology to LG-H. From Lee et al. (2001).

easily identifiable marker that is a good predictor of the presence or absence of a QTL trait can save time and money in a breeding program. Discovery and tagging of QTL is a prerequisite of this type of Marker Assisted Selection (MAS).

The first reported mapping of QTL in soybean was in 1990 (Keim et al., 1990). Since then literally dozens of reports have flowed out of research laboratories with numerous quantitative traits, with perhaps the most critical dealing with seed composition (Diers et al., 1992a), other traits have included reproductive characters, seed characters, maturation traits, vegetative/morphological traits, disease resistance, nutritional efficiency, and more. A thorough coverage of QTL mapping in soybean is too extensive for this chapter. Detailed summarizations of these studies and others can be found in SoyBase, the USDA-sponsored genomic database for soybean, at URL: http://soybase.agron.iastate.edu.

Comparative mapping among grasses has identified highly conserved genome structures and has led to the suggestion that grasses can be considered to have a single genome. This has important implications in our ability to transfer genomic information obtained in one grass species to that of another grass species (Bennetzen and Freeling, 1993). Comparative mapping among legumes has not been as simple. The substantial rearrangements that have occurred within the soybean genome, probably as part of the process of diploidization, make it difficult to identify lengthy stretches of syntenic chromosome segments between soybean and related legumes (Boutin et al., 1995). Although mung bean [*Vigna radiata* (L.) Wilczek, var. radiata] ($2n = 11$) and common bean (*Phaseolus vulgaris* L.) ($2n = 11$) (both belonging to the subtribe Phaseolinae) exhibit a high degree of linkage conservation and preservation of marker order, the situation is substantially different when comparing either one to soybean ($2n = 20$) (subtribe Glycininae). While most linkage groups of mung bean consist of only one or a few linkage blocks from common bean (and vice versa) linkage groups of mung bean and common bean are comprised generally of mosaics of short soybean linkage blocks (Boutin et al., 1995).

A more detailed analysis of homologous segments of soybean, common bean, and mung bean genomes was supportive of the hypothesis that homoeologous chromosome blocks within soybean arose through ancient whole-chromosome duplications (Lee et al., 2001). By focusing on soybean genomic regions containing known duplicated appressed pubescence genes (*Pa1*, *Pa2*), these authors showed that homoeologous segments of soybean linkage groups showed a high degree of synteny with large portions of single chromosomes of *Phaseolus* and *Vigna*. These authors further showed that each of the duplicated and homologous regions among the legumes were homologous with duplicated regions of Arabidopsis.

The first example of inter-family synteny was shown by Grant et al. (2000). Taking advantage of the vast amount of DNA sequence data generated by the Arabidopsis Genome Initiative, Grant et al. (2000) was able to demonstrate synteny between Arabidopsis and soybean. These findings were surprising given the millions of years since the divergence of their lineages. These authors were also able to show that only a limited number of chromosomal events was required to explain the structural differences between soybean and Arabidopsis chromosomes.

For a molecular genetic map to be fully exploited it is necessary to incorporate positions of genes or QTLs in an unambiguous manner. Often, this can be accomplished only through painstaking analysis of segregation data obtained one trait

at a time. However, Shoemaker and Specht (1995) integrated 18 genes into the map in a single experiment. This was done through careful 'stacking' of mutations into parents to be used in constructing the mapping population. Because many of these genes had been put into the classical genetic map, this study resulted in half of the classical genetic linkage groups being integrated into the molecular genetic map in a single experiment. Today, more than 60 loci for qualitative traits have been placed onto molecular maps.

Using an integrated map containing more than 800 markers and combining data from nine different populations, extensive homoeologous relationships were detected using RFLP hybridization techniques (Shoemaker et al., 1996). The average size of these internal duplications is approximately 45 cM, with some duplicated segments covering more than 100 cM. These authors also observed 'nested' duplications that suggested at least one of the original genomes of soybean may have undergone an additional round of tetraploidization in the far distant past (Shoemaker et al., 1996).

6–2 DNA MARKERS AND MOLECULAR GENETIC LINKAGE MAPS

The development of molecular genetic maps based upon DNA sequence polymorphisms was initiated by the suggestion that restriction fragment length polymorphisms (RFLP) could serve as an approach for the development of numerous DNA markers (Botstein et al., 1980). The application of RFLP technology to numerous animal and plant species began shortly thereafter. Subsequently, the availability of the polymerase chain reaction (PCR) (Mullis et al., 1986) as a tool to detect sequence polymorphism led to the development of numerous additional classes of DNA markers. These included (i) microsatellite or simple sequence repeat (SSR) markers (Litt and Luty, 1989; Weber and May, 1989), (ii) random amplified polymorphic DNA (RAPD) (Williams et al., 1990) or arbitrary primer PCR (AP-PCR) markers (Welsh and McClelland, 1990), (iii) DNA amplification fingerprinting (DAF) markers (Caetano-Anolles et al., 1992), and (iv) amplification fragment length polymorphism (AFLP) markers (Vos et al., 1995).

6–2.1 Restriction Fragment Length Polymorphisms-Based Genetic Linkage Maps

The first demonstrations of RFLP in soybean were by Apuya et al. (1988) and Keim et al. (1989) and in 1990 the first RFLP-based map of the soybean genome was published (Keim et al., 1990). To maximize molecular diversity, Keim et al. (1990) constructed their map using a mapping population derived from a cross of cultivated × wild soybean (Table 6–1). This map was developed jointly by the USDA-ARS and Iowa State University with support from the American Soybean Association and saw further expansion during the 1990s with the addition of more than 350 RFLP loci (Shoemaker and Olson, 1993) (Table 6–1). Concurrently, the DuPont corporation (Rafalski and Tingey, 1993) developed an extensive RFLP map with more than 600 loci. Like the USDA-ARS/Iowa State map, Rafalski and Tingey

Table 6–1. Population size and type, linkage group number, map length, and numbers of restriction fragment length polymorphisms (RFLP), simple sequence repeat (SSR), random amplified polymorphic DNA (RAPD), amplification fragment length polymorphism (AFLP), isozyme, and classical genetic markers positioned in published soybean mapping populations.

Mapping population and literature citation	Population size and type	Linkage groups	Total map length	Marker type						
				Total	RFLP	RAPD	SSR	AFLP	Isozyme	Classical
		no.	-cM-	no.						
USDA/Iowa St.										
(A81-356022 × *G. soja* PI 468.916)	59 F2 lines									
Keim et al. (1990)		26	1200	153	150					3
Diers et al. (1992a)		31	2147	246	238				5	3
Shoemaker and Olson (1993)		23	3371	371	355	10			3	3
Shoemaker and Specht (1995)		25	2473	375	358	10			4	3
Cregan et al. (1999)		23	3003	1004	501	10	486	11	4	3
E. I. DuPont Bonus × *G. soja* PI 81762	68 F2 lines									
Rafalski and Tingey (1993)		21	2678	600+	600+					
Univ. of Utah (Minsoy × Noir 1)										
Lark et al. (1993)	69 F3 families	31	1551	140	132				3	5
Mansur et al. (1996)	240 RILs	35	1981	266	224		41		1	6
Cregan et al. (1999)	240 RILs	22	2413	633	209		412		2	10
Univ. of Nebraska (Clark × Harosoy isolines)	57 F2 lines									
Shoemaker and Specht (1995)		26	1004	99	80	3			4	12
Akkaya et al. (1995)		29	1486	123	80	4	34		5	12
Cregan et al. (1999)		28	2787	535	95	69	339	11	7	14
PI 437654 × BSR301	42 RILs									
Keim et al. (1997)		28	3441	840	165	25		650		
Ferreira et al. (2000)		35	3275	356	250	106				
Changnong 4 × Xinnun 6	88 RILs									
Liu et al. (2000)		22	3713	240	100	62	33	42		2
Misuzudaizu × Moshidou Gong 503	190 F2 plants									
Yamanaka et al. (2001)		20	2909	503	401	1	96			5
Kefeng 1 × Nannong 1138-2	201 RILs									
Wu et al. (2001)		24	2321	792						
Noir 1 × BARC-2	149 F2 plants									
Matthews et al. (2001)		35	1400	207	39	17	25	105		4

(1993) relied upon a cultivated × wild soybean cross to increase the relatively low level of RFLP present in cultivated soybean. However, a large proportion of the loci on these two maps would not be expected to segregate in crosses among cultivated soybean genotypes. For example, Shoemaker and Specht (1995) used 358 RFLP markers from the *G. max* × *G. soja* USDA/Iowa State map to genotype progeny derived from a cross of isolines of the cv. Clark and Harosoy. A total of 118 (33%) were polymorphic in the Clark × Harosoy population. In a previous report, Keim et al. (1992) analyzed 38 diverse soybean genotypes with 132 RFLP probes and found 31% to be monomorphic and further, that more than two alleles were detected at only three RFLP loci. In addition to the relatively low level of polymorphism, another complicating factor with the use of RFLP in soybean is the duplicated nature of the soybean genome, to which RFLP probes will hybridize on an average of 2.55 times (Shoemaker et al., 1996). The duplicated nature of the genome results in multiple banding patterns with most RFLP probes. One hybridizing fragment may be mapped in one population and a different or an additional band in another. This requires that an RFLP locus be defined not only by the probe and restriction enzyme being used, but also by the molecular weight of the segregating band(s). Despite these complications, numerous successful analyses designed to discover QTL and characterize genetic variation in soybean germplasm were conducted and reported using RFLP probes from the USDA/Iowa State RFLP map. The details of this and other similar maps and large amounts of related information can be accessed on the World Wide Web in SoyBase, the USDA-ARS Soybean Genome Database (http://soybase.agron.iastate.edu/)

6-2.2 Simple Sequence Repeat Markers

The desire for soybean DNA markers with greater polymorphism was stimulated by the discovery of high levels of allelic variation associated with microsatellite or SSR markers in human (Litt and Luty, 1989; Weber and May, 1989). The fact that SSR markers are PCR based rather than hybridization based was another attractive feature of this DNA marker system. In the early 1990s, two research groups published similar reports demonstrating the high levels of polymorphism, co-dominant inheritance, and the locus specificity of SSR markers in soybean (Akkaya et al., 1992; Morgante and Olivieri, 1993). Akkaya et al. (1992) found as many as eight SSR alleles at one locus in a set of 38 *G. max* and five *G. soja* genotypes. Subsequent reports of SSR allelic variation in cultivated and wild soybean (Cregan et al., 1994; Maughan et al., 1995; Morgante et al., 1994; Rongwen et al., 1995) detected very high levels of allelic variation, including one locus with 26 alleles among a group of 91 cultivated and five wild soybean genotypes. In addition, data analyses suggested little evidence of the clustering of SSR loci in the soybean genome (Akkaya et al., 1995). Because of their high levels of polymorphism, single locus nature, and random distribution in the genome, it was concluded that SSR markers would provide an excellent complement to RFLP markers for use in soybean molecular biology, genetics, and plant-breeding research. The major drawback to SSR markers is the high cost of development, which requires firstly the discovery of SSR motifs and secondly, knowledge of the flanking sequence to permit the design of locus-specific PCR primers. An additional technical difficulty associated

with SSR technology is the frequent need to distinguish alleles that vary by only one or a few repeat units in size.

6–2.3 Restriction Fragment Length Polymorphisms and DNA Amplification Fingerprinting Markers

In contrast to SSRs, RAPD or AP-PCR markers require no prior knowledge of DNA sequence and as a dominant marker, alternative alleles are detected simply as the presence or absence of a PCR product. Thus, genotypes can be readily determined using agarose gel electrophoresis without the need for more sophisticated systems to detect allelic variation. Nonetheless, RAPDs have not been widely used in soybean genetic map development. The exception is the RFLP/RAPD map constructed by Ferreira et al. (2000), which incorporated 106 RAPD markers into a framework of 250 existing RFLP loci using a subset of RILs from the PI 437654 × BSR 101 population (Table 6–1). Like RAPD markers, DNA amplification fingerprint or DAF markers are amplified using a single arbitrary primer (Caetano-Anolles et al., 1992). The differences between RAPD and DAF technology are the shorter arbitrary primer in DAF vs. RAPD (generally eight nucleotides), and the use of polyacrylamide gel electrophoresis with silver staining in the case of DAF markers vs. agarose gels for RAPDs. Predigestion of genomic DNA with a restriction enzyme before PCR amplification is sometimes used to optimize DAF amplification products. A limited number of DAF-generated polymorphisms were mapped in the Univ. of Utah, Minsoy × Noir 1 RIL population (Prabhu and Gresshoff, 1994). No genetic maps were developed in soybean using DAF markers.

6–2.4 Amplification Fragment Length Polymorphism Markers

Like RAPD markers, the generation of AFLP requires no prior knowledge of DNA sequence and as a result, numerous marker loci can be rapidly developed. The AFLP markers are generated based on restriction fragment length polymorphisms. The DNA adaptors are ligated to the ends of restriction fragments and PCR primers homologous to the adaptors are used to amplify selected subpopulations of the pool of fragments. Selectivity results from the addition of two or three arbitrary nucleotides to the 3′ ends of the PCR primers. One of the largest available AFLP maps of any plant species was developed in soybean (Keim et al., 1997). These loci were mapped in a subset of 42 RILs from the 330 RIL PI 437654 × BSR 101 population. This map has a total of 840 loci, 650 of which are AFLPs, whereas the USDA/Iowa State map has 1004 loci, most of which are RFLPs and SSRs (Table 6–1). The authors noted significant clustering with AFLP markers that were generated using *Eco*RI/*Mse*I restriction enzymes (Young et al., 1999). Thirty-four percent of the loci displayed dense clustering. In contrast, *Pst*I/*Mse*I-generated AFLP loci did not cluster. *Pst*I is known to be sensitive to cytosine methylation and in relation to *Eco*RI/*Mse*I-generated AFLPs, those produced using *Pst*I/*Mse*I appeared to eliminate marker clustering. Keim et al. (1997) also noted that despite numerous marker loci and the framework of 165 RFLP loci, 11 of 28 linkage groups could not be aligned with a homologous linkage group on the USDA/Iowa State map. This result is indicative of one shortcoming of AFLP markers, which is the difficulty of

comparing AFLP loci across populations. As a result, it is generally necessary to create a new linkage map for every new population, rather than simply using a set of informative marker loci with previously defined positions in the genome. However, because of the large amount of marker data that can be obtained with AFLP without the need for previous knowledge of DNA sequence, AFLPs have proven useful for saturating specific genomic regions using bulked segregant analysis (Michelmore et al., 1991) or for the comparison of near-isogenic lines for a trait of interest (Muehlbauer et al., 1988, 1991). The loci so identified can then be associated with an anchor marker to determine a definitive position on a linkage map.

6–2.5 A Simple Sequence Repeat-Based Soybean Genome Map

The development and mapping of a large set of soybean SSR markers was initiated in 1995 with the support of the United Soybean Board. As a result of this effort more than 600 SSR loci were developed. These markers were derived almost exclusively from genomic libraries. Clearly, the large collection of soybean EST data is another source of sequence data that could be exploited for the development of additional SSR markers. The virtue of EST-derived SSRs is the close association of the resulting SSR markers with expressed genes. The 600 SSR markers developed to date were mapped in one, two, or when possible, three different mapping populations (Cregan et al., 1999). One of these was the USDA/Iowa State population and the second was the 240 RIL University of Utah population developed from a cross of the cultivated soybean genotypes Minsoy and Noir 1 (Table 6–1). The third was the University of Nebraska Clark × Harosoy isoline population consisting of 57 F2-derived lines. In every case, SSR markers mapped to a single locus in the genome with a map order that was essentially identical in all three populations. A total of 187 loci were mapped in each of the three populations, making it a simple matter to align homologous linkage groups. An example of this alignment is presented in Fig 6–2, which pictures one of the 20 homologous soybean linkage groups. In this case, 12 loci were mapped in all three populations and an additional 20 in two populations. Thus, linkage groups U13a and U13b on the Univ. of Utah map could be joined based upon markers in common with linkage groups F-ISU and F-CH21 from the USDA/Iowa State Univ. and the Univ. of Nebraska maps, respectively. As a result, the 20+ linkage groups derived from each of the three populations (Table 6–1) were aligned into a consensus set of 20 homologous groups presumed to correspond to the 20 pairs of soybean chromosomes. Likewise, classical linkage group 8 (*Ms1*, *W1*, *Ms6*, *Y23*, *St5*, and *Adh1* loci) was associated with linkage group F by in situ segregation of *W1* as were classical loci *Rpg*1 and *B1* that had not previously been linked to any other classical loci. Reports in the literature indicated that both *Rps*3 (Diers et al., 1992b) and *Rsv*1 (Yu et al., 1994) mapped to linkage group F, thereby associating classical linkage group 13 with linkage group F. Based upon in situ segregation or linkage reports in the literature, all but one of the 20 classical linkage groups (Palmer and Shoemaker, 1998) were assigned to a corresponding molecular linkage group. Work is being completed to position approximately 100 additional classical loci on the integrated map using bulked hybrid segregation (Rector et al., 1999). Information relating to the sequence of PCR primers to the 600+ SSR loci reported in Cregan et al. (1999) and a standard pro-

tocol for their amplification can be obtained on the SoyBase website. Additional information relating to SSR allele sizes in a set of 10 diverse soybean genotypes, as well as gel images of the alleles produced with the same 10 genotypes, is available on SoyBase. Data relating to the mapping of more than 600 SSR in the University of Utah Minsoy × Noir 1, Minsoy × Archer, and Archer × Noir recombinant inbred line populations can be obtained on the Lark Lab, Univ. of Utah website (http://www.larklab.4biz.net/)

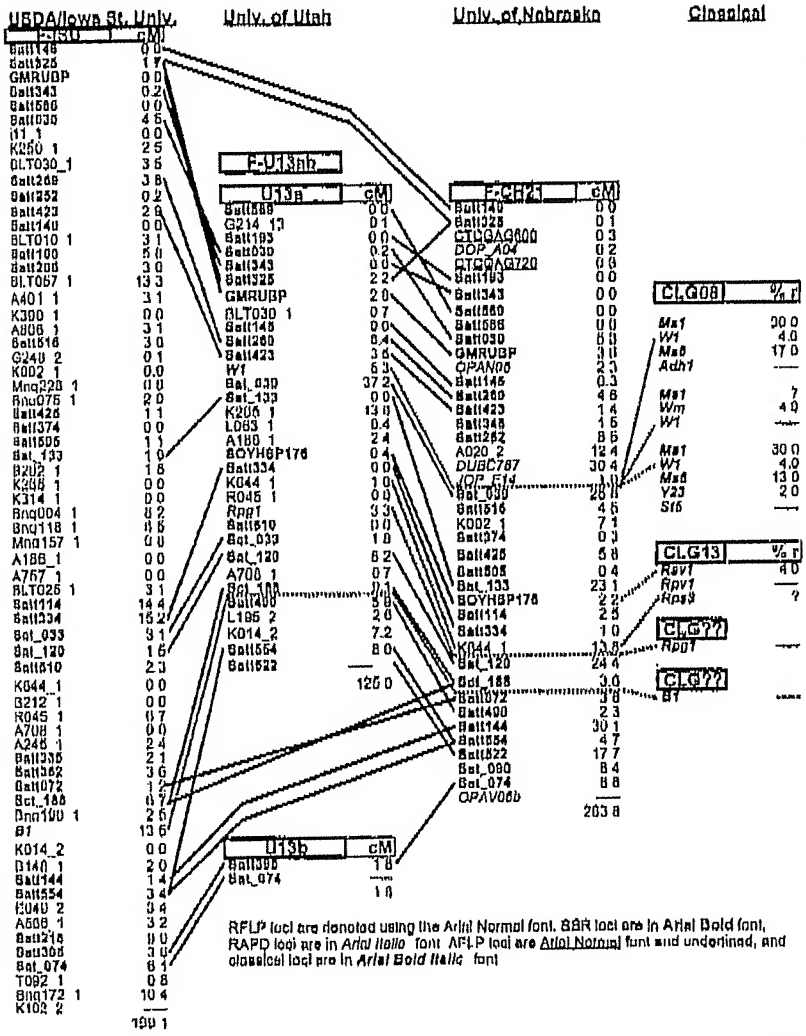

Fig. 6–2 Consensus soybean molecular linkage group F defined using three mapping populations: The USDA/Iowa State Univ., *Glycine max* × *G. soja* population; the Univ. of Utah, Minsoy × Noir 1 population; and the Univ. of Nebraska, Clark × Harosoy population. From Cregan et al. (1999).

The definition of 20 consensus linkage groups with an average of 30 locus-specific SSR markers per group (Cregan et al., 1999) provided a resource that has facilitated the rapid alignment of linkage groups in existing or newly created linkage maps with the consensus linkage groups in the SSR-based genome map. In four instances, a relatively small number of SSR markers ranging from an average of one to as many as five per linkage group was used to associate linkage groups with corresponding linkage groups on the SSR-based soybean genome map. These included two fairly extensive maps, one with 792 markers (Wu et al., 2001) and another with more than 500 markers (Yamanaka et al., 2001) (Table 6–1). Details of the Misuzudaizu × Moshidou Gong 503 map (Yamanaka et al., 2001) are available on the World Wide Web (http://dna-res.kazusa.or.jp/8/2/02/HTMLA/). Matthews et al. (2001) successfully positioned cDNA and genomic clones on consensus linkage groups in the SSR-based soybean genome map using a small number of SSR loci to align corresponding linkage groups (Table 6–1). Liu et al. (2000) provided the fourth example of the use of SSRs to align linkage groups with the consensus map.

6–2.6 Single Nucleotide Polymorphism Markers

Single DNA base changes between homologous DNA fragments plus small insertions and deletions, collectively referred to as single nucleotide polymorphisms (SNPs), are by far the most abundant source of DNA polymorphisms in humans (Collins et al., 1998; Kruglyak 1997; Kwok et al., 1996) and mice (*Mus musculus*) (Lindblad-Toh et al., 2000). In humans, these variations are estimated to occur at a frequency of about one per 1000 bp when any two homologous DNA segments are compared (Cooper et al., 1985; Kwok et al., 1996). In plants, relatively limited data on the frequency of SNPs are available. Cho et al. (1999) compared the DNA sequence of more than 500 kbp of the two *Arabidopsis thaliana* genotypes Columbia and Landsberg erecta and detected one SNP every 1034 bp. However, most other reports have indicated much higher levels of sequence variation in *Arabidopsis* (Kawabe et al., 2000; Kawabe and Miyashita, 1999; Kuittinen and Aguade, 2000; Purugganan and Suddith, 1999). In maize (*Z. mays* ssp. *mays* L.), Tenaillon et al. (2001) sequenced more than 14 kb of coding and noncoding DNA from 21 loci on chromosome 1 in each of 25 genotypes and discovered a mean of 9.6 SNPs per kbp between any two randomly selected genotypes. In soybean, SNP DNA markers are already in use in industrial-breeding programs (Cahill, 2000) using allele-specific hybridization (ASH) for SNP detection similar to the procedure described by Coryell et al. (1999). It is thus apparent that SNP markers are likely to have an important role in the future of soybean genome analysis and manipulation.

Until recently, the comparison of variation in DNA sequence among soybean genotypes has been confined to the assay of single genes or DNA fragments, generally with the purpose of defining gene structure or function or evolutionary relationships. For example, Scallon et al. (1987) compared 3543 bp of the Gy_4 glycinin gene plus flanking DNA in two genotypes and found three SNPs. Zakharova et al. (1989) compared 789 bp of cDNA sequence encoding the A_3B_4 glycinin subunit of the soybean cv. Mandarin, Ramnaya-10, and Mukden and found two single nucleotide polymorphisms. Xue et al. (1992) discovered 20 single base changes and

four indels (insertion-deletions) in a comparison of 2942 bp of the *Gy4* gene and flanking DNA in the soybean genotypes Forrest, Raiden, and Dare. Zhu et al. (1995) sequenced 400 bp of RFLP probe A-199a in the cv. BSR-101 and A81-356022 and the *G. max* germplasm line PI437654 and found a total of nine SNPs. To permit the comparison of SNP frequency among loci of varying length and between populations that vary in size, measures of nucleotide diversity such as π (Tajima, 1983) and Watterson's theta (θ_w) (Watterson, 1975) have been devised that are standardized for length and adjusted for sample size. Nucleotide diversity from the four aforementioned studies range from $\theta_w = 0.85$ SNPs/kbp (Scallon et al., 1987) to $\theta_w = 15$ SNPs/kbp (Zhu et al., 1995). The wide diversity of values suggested that a systematic study of SNP frequency in soybean was needed.

In recently completed work to assess the SNP frequency in soybean, a group of 25 soybean genotypes that represented 18 ancestral varieties from which North American soybean plants are derived (Gizlice et al., 1994) as well as seven parents of RIL mapping populations was analyzed (Zhu et al., 2003). A total of more than 28.5 kbp of coding sequence and 37.9 kb of noncoding (introns, 3′ and 5′ UTR, and flanking genomic sequence) from 116 genes was sequenced in each of the 25 genotypes. The SNP frequency in coding and noncoding DNA was 1.98 kbp^{-1} and 4.19 kbp^{-1}, respectively. Nucleotide diversity was $\theta_w = 0.53$ and 1.11 in coding and noncoding sequence, respectively. The mean $\theta_w = 0.97$ was similar to reports of SNP frequency in humans (Wang et al., 1998; Cargill et al., 1999; Halushka et al., 1999) and 5- to 10-fold lower than reports in maize (Remington et al., 2001; Tenaillon et al., 2001). Despite the relatively low frequency, SNPs were discovered in or around 74 of the 116 genes for which sequence data were obtained. These data suggested that SNP discovery focused on noncoding sequence, where greater sequence polymorphism is present, will permit successful SNP discovery in soybean. One obvious target for SNP discovery is 3′ UTRs of cDNAs. Discovery and mapping of SNPs in 3′ UTRs will not only create useful genetic markers but will position the corresponding expressed gene on the genetic map. The resulting transcript map will provide a powerful tool to associate QTL with candidate genes. An alternative approach to SNP discovery is the sequence analysis of polymorphic AFLP bands or adjacent polymorphic sites as suggested by Meksem et al. (2001). This approach was successful in discovering numerous indels and SNPs in soybean.

6–3 TECHNOLOGIES FOR DNA MARKER ANALYSIS

6–3.1 Restriction Fragment Length Polymorphisms and Random Amplified Polymorphic DNA

The electrophoretic separation of DNA fragments, transfer and immobilization on a membrane, and detection of specific sequences was outlined by Southern (1975) and is the basis for the detection of RFLP. Numerous descriptions of the procedure are available (Sambrook et al., 1989; Grant and Shoemaker, 1997) and as a result there is little need for detailed description here. Likewise, there is little need to describe the analysis of RAPD markers. As indicated earlier, two of the important virtues of RAPD markers are ease of use and broad applicability. Almost

without exception, RAPD fragments are analyzed using agarose gel electrophoresis (Rafalski, 1997). High resolution agaroses such as Metaphor agarose (FMC Bioproducts), Agarose 3:1 (Amresco, Solon, OH), and Synergel (Diversified Biotech, Inc., Boston, MA) are also used for RAPD fragment analysis.

6–3.2 Amplification Fragment Length Polymorphism

The procedures that Vos et al. (1995) outlined for the detection of AFLP used standard sequencing gels for the analysis of ^{32}P end-labeled AFLP fragments. In developing the soybean AFLP-based map, Keim et al. (1997) used a slightly modified AFLP protocol as described by Travis et al. (1996). Techniques for chemiluminescent detection of AFLP fragments are available (Lin et al., 1999). The AFLP technologies have been modified to function with small slab gels, a discontinuous buffer system, and silver staining (Mano et al., 2001). The AFLP fragment sizing has also been adapted to numerous semi-automated analysis platforms including the ABI PRISM 377 DNA and the Licor Global IR^2 DNA Analyzer.

6–3.3 Simple Sequence Repeat

Because SSR alleles are defined by the number of repeat units present in the SSR, the basis of allele sizing is the determination of PCR product size. As a result, numerous methods can be used for allele discrimination. With the exception of time-of-flight mass spectrometry, these procedures depend upon analyses using either slab gel or capillary electrophoresis. A necessarily brief overview of available options for SSR allele sizing follows.

6–3.4 Agarose Gel Electrophoresis

In those instances when allele sizes vary by six to eight or more basepairs, high resolution agarose gels are frequently quite adequate for purposes of mapping and marker assisted selection. Appropriate high-resolution agaroses are listed in the preceding paragraph.

6–3.5 Polyacrylamide Gel Electrophoresis

Initially, most SSR allele sizing was performed on high-resolution denaturing polyacrylamide sequencing gels to distinguish fragments that might differ in length by only one or two bases. This standard procedure was described by Kraft et al. (1988) and is used to separate products that are either internally labeled or end-labeled with ^{32}P or ^{33}P. Frequently, numerous "shadow bands" or "stutter bands" are associated with an SSR-containing fragment, making accurate size determination difficult, especially in the case of dinucleotide repeats. Some reduction in stutter bands can be achieved by using sequencing gels containing formamide in addition to urea as suggested by Litt et al. (1993). Silver staining is also frequently used to visualize SSR alleles as applied by Mansur et al. (1996). To resolve SSR alleles that differ by three or more bases, Cregan and Quigley (1997) suggested SSR allele sizing on 1.5-mm thick denaturing formamide/urea gels followed by stain-

ing with the DNA-specific stains SYBR-green or SYBR-gold (FMC Bioproducts) and detection on a UV transilluminator. Numerous other gel and capillary electrophoresis systems are available for high throughput automated or semi-automated SSR allele sizing. These include gel and capillary electrophoresis systems from Applied Biosystems (Foster City, CA), LI-COR, Inc. (Lincoln, NE), Beckman Coulter (Fullerton, CA) and Amersham Biosciences (Piscataway, NJ). Advantages of these systems over standard sequencing gels for sizing SSR-containing PCR products are (i) single-base resolution over a wide size range from 75 to 500 bases, (ii) automated sizing, (iii) automated data output, and (iv) elimination of radioactivity. Numerous soybean researchers have reported the use of the ABI PRISM 377 DNA for SSR allele sizing in soybean (Diwan and Cregan, 1997; Mian et al., 1999; Song et al., 1999; Narvel et al., 2000a, 2000b).

6–3.6 Capillary Electrophoresis

The same fluorescent chemistry employed in the Perkin-Elmer DNA sequencers described above is used in capillary electrophoresis systems with 1, 16, or 96 capillary capacity available from Perkin-Elmer Applied Biosystems. Beckman Coulter, Inc. manufactures an eight capillary machine and Amersham Biosciences has three systems with 48, 96, or 384 capillaries that can be used for multiplex SSR allele sizing.

6–3.7 Mass Spectrometry

Braun et al. (1997) proposed the use of matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF) for SSR allele sizing. This system requires the annealing of a single detection primer within a few bases of the 3′-end of the SSR. A DNA polymerase extends this primer through the SSR by primer-directed DNA synthesis. A dideoxynucleotide triphosphate (ddNTP) is included to terminate the reaction at a point past the 5′-end of the SSR. Extension reactions from different SSR alleles yield products that differ in length by the number of bases in the alleles. These products are resolved using MALDI-TOF mass spectrometry.

6–3.8 Single Nucleotide Polymorphisms

The promise of SNP markers is the efficiencies in cost per data point and speed of data acquisition that will result from the technological innovations likely to be forthcoming from intensive research that has and continues to be focused on SNP detection. Numerous reviews of these technologies are available (Brookes, 1999; Gut, 2001; Kwok, 2000; Kwok and Chen, 1998; Shi, 2001; Syvanen, 2001). There are four basic approaches to SNP detection. These include (i) allele-specific hybridization (ASH) or allele-specific oligonucleotide hybridization, (ii) single-base extension (SBE) or minisequencing, (iii) the oligonucleotide ligation assay (OLA), and (iv) allele-specific cleavage of a "flap probe". These approaches have been combined with numerous different detection technologies. A number of these are summarized in Table 6–2. In many cases a SNP-containing PCR product is the neces-

Table 6–2. A partial listing of technologies used in the detection of single nucleotide polymorphisms (SNPs).

Allele specific hybridization	
^{33}P-labelled probe hybridized to a SNP-containing fragment immobilized on a membrane	Coryell et al. (1999)
5′ nuclease assay	Lee et al (1993)
Molecular beacons	Tyagi et al. (1998)
Electronic dot blot on semiconductor microchips	Gilles et al. (1999)
Electric field denaturation	Sosnowski et al. (1997)
Affymetrix oligo chip	Sapolsky et al. (1999)
Massecodes cleaved from allele specific oligos detected via mass spectrometry	Kokoris et al. (2000)
Randomly ordered fiber-optic gene arrays	Steemers et al. (2000)
Flow cytometry	
eSensor™	Yu et al (2001)
Dynamic allele-specific hybridization (DASH)	Prince et al (2001)
Single-base extension or minisequencing	
Single base extension-Tag array on glass slides (SBE-TAGS)	Hirschhorn et al. (2000)
Matrix-assisted laser desorption ionization-Time of flight (MALDI-TOF) mass spectrometry	Little et al. (1997); Ross et al. (1998)
Fluorescent dideoxynucleotide triphosphates (ddNTPs)	Lindblad-Toh et al (2000)
Flow cytometry	Chen et al. (2000)
Pyrosequencing (multiple base extension)	Alderborn et al. (2000)
Denaturing high performance liquid chromatography	Hoogendoorn et al. (1999)
Fluorescence polarization	Chen et al. (1999)
Oligonucleotide ligation assay	
Rolling circle amplification	Qi et al. (2001)
Flow cytometry	Iannone et al. (2000)
Allele-specific cleavage of a "flap probe"	Fors et al. (2000)

sary target for detection; however, in some instances genomic DNA is the target and a PCR step is not required. As of December 2002, nearly 5 million putative human SNPs had been submitted to dbSNP (http://www.ncbi.nlm.nih.gov/SNP/index.html), the National Center for Biotechnology Information, National Institutes of Health, SNP database. Clearly, the human genetics community is focusing on SNPs as a major research tool for drug discovery, diagnostics, gene discovery, population genetics, and other applications. The costs of SNP detection are likely to decrease while the ease and speed of detection improve. The plant genetics community stands to be a beneficiary of the investments in technology being made by human geneticists.

6–4 GENE DISCOVERY

6–4.1 Expressed Sequence Tags

For nearly two decades random sequencing of gene transcripts has been recognized as a simple and efficient method of identification of many of the expressed genes in an organism (Putney et al., 1983). These sequences, known as Expressed Sequence Tags (ESTs), have become a valuable and efficient method for gene discovery (Sterky et al., 1998; Hillier et al., 1996; Marra et al., 1999). When sampling is random, the frequency of appearance of any given EST permits the identifica-

tion of differential patterns of gene expression (Manger et al., 1998; Tanabe et al., 1999, Ewing et al., 1999). The ESTs also provide an opportunity to study gene evolution, to make comparative analyses between genera, and coupled with genetic mapping can identify candidates for important biological processes (Hatey et al., 1998).

Global, multi-tissue EST projects have been reported for *Arabidopsis* (Delseny et al., 1997) and rice (Ewing et al., 1999). More specialized, tissue-specific EST projects have been reported from root-hair-enriched *Medicago truncatula* tissue (Covitz et al., 1998), flower buds of Chinese cabbage [*Brassica rapa* (Pekinensis Group)](Lim et al., 1996), and wood-forming tissues of poplar (*Populus* spp.) (Sterky et al., 1998).

Shoemaker et al. (2002) reported on a global, multi-tissue EST analysis for soybean. More than 120 000 ESTs were generated from more than 50 cDNA libraries representing a wide range of organs, developmental stages, genotypes, and environmental conditions. This study was able to demonstrate correlated patterns of gene expression across cDNA libraries. As a result, gene expression profiles could be evaluated across libraries, cDNA libraries with similar EST composition and genes with similar expression patterns and potentially similar functions were grouped. These studies provide a large resource of publicly available genes and gene sequences and provide valuable insight into structure, function, and evolution of a model crop legume.

6-4.2 Genome Sequencing

Genome sequencing is the cornerstone of functional analyses and is fundamental to understanding the genetic composition of an organism. Whole-genome sequencing is currently underway for rice and is complete for all five of the *Arabidopsis* chromosomes, with gap-filling and annotation in progress (Theologis et al., 2000; Salanoubat et al., 2000; Tabata et al., 2000; Mayer et al., 1999; Lin et al., 1999). Because of the size and complexity of the soybean genome, it is unlikely that, given the current technology, the entire genome will be sequenced in the near future.

'Genomic Sampling' of nearly 2700 DNA sequences from more than 600 mapped loci has provided a glimpse of the composition and general structure of the soybean genome (Marek et al., 2001). For example, approximately one-third of all sequences sampled near SSR markers corresponded to repetitive DNA, while a little more than 18% of the sequences sampled near RFLP markers (putative hypomethylated regions) corresponded to repetitive DNA. Surprisingly, about 7 to 10% of the sequences sampled had significant similarity to an existing soybean EST sequence (Marek et al., 2001). Additionally, the clustering of BAC-end sequences around an anchored locus provided the opportunity to look for micro-synteny between soybean and other species. Few examples of microsynteny were observed between soybean and *Arabidopsis*, while about 33% of the sequence clusters detected microsynteny between soybean and *Medicago truncatula*, another legume (Marek et al., 2001).

Other 'sampling' approaches such as sequencing of hypomethylated regions or selection for open-reading frames (ORFs) may provide a wealth of information about gene-rich regions without the expense of whole-genome sequencing. Ap-

proaches that select for genes whose expression varies by organ and tissues, and during various stages of development or under different environmental and biological stresses can also provide a wealth of information. These combinations of data provide essential information about the regulation of genes and about metabolic regulation of the organism.

6–5 FUNCTIONAL GENOMICS

Breakthrough technologies stimulated by the human genome project are fundamentally changing the way in which biology is conducted in the genomics and postgenomics era. The approach has shifted from experimental analysis of "one gene at a time" to "thousands at a time" (with microarrays and chips), from "one protein at the time" to "thousands at a time" (proteomics), and to whole metabolite profiling (metabolomics) of cells and tissues (Brent, 2000; Lander and Weinberg, 2000).

In the broadest view, functional genomics is defined as the process of generating, integrating, and using information from genomics (sequencing), gene expression profiling (microarrays and chips), proteomics, metabolic profiling, and large-scale genotyping and trait analysis to understand the function of genes. Bioinformatics, statistical sciences, and computational sciences are increasingly indispensable for functional genomics research as the field becomes more driven by large-scale data and information processing (Heiter and Boguski, 1997).

The new Plant Genome Program, which received $85 million of funding during 1998 from the National Science Foundation (NSF), stimulated the development of structural and functional genomics resources for plants other than the model plant *Arabidopsis*. The program funded collaborative, multidisciplinary research in maize, soybean, tomato, cotton (*Gossypium hirsutum* L.), and other plants of agronomic importance to speed the development of tools that would be publicly available for gene expression analysis, gene tagging, and mapping (Walbot, 1999).

6–5.1 Creating a Soybean "Unigene" Set

One goal of the NSF-sponsored program for "Soybean Functional Genomics" is to develop a set of 30 000 unique genes from soybean. To accomplish this, the EST data from the commodity-sponsored "Public EST Project" (Shoemaker et al., 2002) is used as the raw material. The mRNAs that are more abundant in various tissues will be more highly represented in the EST collections. The ESTs are compared by computer programs such as PHRAP (Green, 2001) to assemble them into overlapping clones that have sequence similarity known as contiguous segments (contigs). In this way, longer sequences representing expressed genes are assembled and identical sequences that represent redundant clones are recognized. For example, contigs representing the storage protein Kunitz trypsin inhibitor consist of 118 ESTs that form a contig from among 2600 sequences in a cDNA library from the developing cotyledons of soybean whereas the contig representing soybean lipoxygenase-2 consists of six member sequences. The number of sequences in the

contigs in a non-normalized cDNA library is a rough approximation of the relative abundance of the mRNAs within that tissue.

To develop a "unigene" set for soybean, the EST representing the most 5′ sequence read in each contig is selected to represent that particular gene as it will likely represent the longest clone. In addition, the many sequences that occur only once in the EST collection (singletons) are also selected. Thus, the combined number of singletons and independent contigs that result from a computer assembly of ESTs becomes an estimate of the number of unique genes in the organism. This process is continuous because as the number of ESTs grows, the number of unique genes in the organism will continue to be refined. To date, this number for soybean is exceeding 40 000 (Shoemaker et al., 2002; Vodkin et al., 2003) from a collection of more than 208 000 ESTs. The *Arabidopsis* genome sequence has revealed approximately 26 000 genes (Arabidopsis Genome Initiative, 2000; Martienssen and McCombie, 2001). The number of human genes has been estimated at approximately 30 000 sequences from the complete human genome sequence (International Human Genome Sequencing Consortium, 2001); however, that number is still under debate (Hegenesch, et al., 2001). Alternative splicing appears to result in a much larger number of proteins in the human transcriptome than previously thought. The degree of alternative splicing in soybean or other higher plants has not been assessed. However, many higher plants, including soybean, have extensive duplications of sequences. Many functionally equivalent proteins are encoded by small gene families. One of the surprises to come from the knowledge of the complete *Arabidopsis* genome sequence was the extent to which the small genome of this plant contained duplicated sequences (Marteinssen and McCombie, 2001).

6–5.2 Global Gene Expression Analysis

The genomes of higher eukaryotes contain a very large number of genes. Depending upon the organism, this predicted number ranges from 20 000 to >60 000 unique genes. All of these genes are expressed in a coordinated fashion across tissues, across developmental time, and in response to particular physiological conditions. This regulation may be tightly linked for suites of genes (e.g., a particular biochemical pathway) and some transcripts will alter patterns of expression for other genes. Traditionally, single gene expression patterns have been studied to understand how different temporal, developmental, and physiological processes affected gene expression. This has resulted in our current models for gene regulation but has limited our ability to understand complex regulatory relationships among genes. With recent advances in genomics, very large numbers of genes can now be simultaneously analyzed for their expression levels in a comparative fashion between two biological states using microarray or biochip technology.

Several techniques for high throughput or global analysis of gene expression have been described as an outgrowth of the human genome project (Velculescu et al., 1995; Schena et al., 1995, 1996; DeRisi et al., 1997; Marshall and Hodgson, 1998). These include (i) high density expression arrays of cDNAs on conventional nylon filters with radioactive probing, (ii) microarrays or "chips" using fluorescent probes, and (iii) serial analysis of gene expression (SAGE).

6–5.2.1 High Density cDNA Arrays with Radioactive Probing

High density arrays are one method for assessing gene expression. The cDNA clones are arrayed in 384-well plates and spotted by robots onto nylon membranes. Either the bacterial colony, plasmid DNAs, or PCR products can be spotted. As many as 18 000 cDNAs can be arrayed on a filter of about 20 by 20 cm using robotic technology. The hybridization method is analogous to that used for conventional DNA or RNA blots on nitrocellulose membranes. The membranes are generally probed with ^{33}P-cDNA label produced by reverse transcription of mRNA. After hybridization, the membrane is imaged using a phosphorimager and the highly complex pattern must be read by image analysis software. This technology can be used to select specific cDNAs that are very weakly expressed in the library for further analysis. This "filter normalization" method is a tool for gene discovery of more weakly expressed genes. High density expression arrays have been used to increase gene discovery in soybean (Vodkin et al., 2000). In addition, one can compare the expression within a single library of numerous genes. If a "unigene" collection is used, then the relative expression of various genes within a single sample is easily obtained from high density membranes. The disadvantage is that without dual labeling, one cannot easily compare between two different mRNA samples. Dual labeling is one of the main advantages of microarrays using fluorescent detection.

6–5.2.2 DNA Microarrays or Chips to Analyze Global Gene Expression Patterns

An alternative method to the high density filters is microarray technology using fluorescent probes (DiRisi et al., 1997; Schena et al., 1995). In this method, the inserts from cDNA clones are amplified by PCR with vector primers and the amplified DNAs are arrayed onto glass microscope slides by a computer-controlled printing device. After fixation to the slide, the DNAs on the array are probed with fluorescently labeled cDNAs made from total mRNA of a particular tissue. Two different fluorescently labeled probes can be used simultaneously on the same slide, that is, one in the red range and one in the green range. The fluorescent images are captured with a scanning laser microscope and the intensity of each spot can be compared to standards of known concentrations to give quantitative data on gene expression. An arrayed set of 2375 unique cDNA from *Arabidopsis* has been used to examine changes in gene expression during pathogen challenge (Schenk et al., 2000) and will be used to address important problems in many other plant systems (Bouchez and Hofte, 1998; Mazur et al., 1999; DellaPenna, 1999; Somerville and Somerville, 1999; Somerville, 2000).

Microarrays are most effective when all genes within an organism are represented. Current technology for cDNA arrays are typically in the range of 5000 to 10 000 genes on glass slides. Thus, it is important to reduce the redundancy of the cDNA libraries before they are spotted to maximize the number of genes on the array by printing from the "unigene" set instead of from non-normalized libraries. Currently, arrays of 27 000 cDNAs (9k per array) for soybean have been printed containing low redundancy cDNA from many of the 80 cDNA libraries of the EST project. Experiments to examine differential expression during the process of induction

of somatic embryos from tissue culture have been conducted (Thibaud-Nissen et al., 2003). The applications of microarray technology to soybean are enormous. A few include profiling the genes that respond to challenges by various pathogens and environmental stresses such as drought, heat, cold, flooding, and herbicide application. In addition, expression profiling of isoline genotypes that differ in protein or oil content and other quantitative traits will yield significant clues to the genes involved in those pathways and traits.

The future of functional genomics research will include oligonucleotide-based glass arrays that will allow distinguishing gene family members. Full-length sequencing of cDNA clones will include information at the 5′ and 3′ untranslated regions. The 3′ region has more sequence variability and can be used to design oligonucleotides that distinguish gene family members. Assembly of the full-length cDNAs will also allow prediction of the entire ORF. Affymetrix chip technology in which approximately 20 nucleotides are synthesized directly using a photolithographic mask (Marshall and Hodgson, 1998) is very expensive and also requires full-length sequence of the expressed soybean genes. Full-length cDNA information is also critical for large-scale proteomics research on soybean so that peptide masses or partial protein sequences can be matched to the predicted ORFs. Although gene knockout and transposon tagging are not easy to obtain in soybean on a global scale, the application of microarrays for gene expression profiling and proteomics to investigate natural genetic variation in soybean promises to yield substantial information over the next decade that will aid in determining the function of soybean genes.

6–5.2.3 Serial Analysis of Gene Expression

Serial analysis of gene expression represents both a qualitative and quantitative method to characterize gene expression and to compare these patterns across tissues (Velculescu et al., 1995). SAGE captures short 10 to 20 nucleotide "tags" near the 3′ end of individual mRNA molecules. It has been shown that 10 nucleotides uniquely identify >95% of human EST sequences. The tags are ligated into concatamers and are then cloned into plasmid vectors for sequencing (~40 tags per clone). Hundreds or even thousands of clones are sequenced to identify SAGE tags and to determine frequency in the plasmid library. The frequency of appearance of tags in the library has been shown to accurately estimate expression levels in the mRNA source tissue. Messages at a very low expression level (one per cell) can be quantitatively detected by sequencing numerous plasmid clones. More highly and moderately expressed genes are easily detected with only a few hundred sequenced clones. SAGE analysis requires an extensive 3′ sequence data base. A disadvantage is that low abundance tags are not detected without extensive sequencing of the SAGE tag library from that source, which can be expensive. Initial analysis from 20 SAGE libraries in soybean has resulted in 132 992 SAGE tags of which 40 121 are unique (J. Schupp and P. Keim, personal communication, 2003).

6–5.3 Functional Consequences of Gene Duplication

Soybean has many gene family members that consist of 2 to 10 members that are very similar in sequence (Graham et al., 2000). In addition to gene duplications

that arose because of the ancient tetraploid nature of soybean, there are more recent duplications that give rise to gene family members. For example, the chalcone synthase (CHS) family in soybean consists of six gene family members that are >97% similar in the protein-coding region (Akada and Dube, 1995). Five of these reside on the same BAC clone and one is unlinked. Gene family members contain more variation in the 5′ and 3′ untranslated regions that allow for gene-specific primers that can be used for experiments to detect differential expression (Todd and Vodkin, 1996).

In addition to providing the raw material for evolution, gene duplications sometimes lead to gene silencing or cosuppression. Cosuppression is a phenomenon, first described in plant systems in the early 1990s, whereby additional copies of a CHS gene in transgenic petunia (*Petunia hybrida* Vilm.) leads to suppression of both the endogenous and transgenic CHS transcripts (reviewed in Jorgensen, 1995). Homology-dependent gene silencing is now recognized widely in eukaryotic systems (Wolffe and Matzke, 1999). In soybean, the *I* (inhibitor) locus is a classical dominant genetic marker that results in yellow seed coats. The dominant form of the locus is present in most commercially used soybean varieties but spontaneous mutations to the recessive *i* allele occur and result in pigmented seed coats. Total levels of the CHS mRNAs and enzyme activity are reduced in yellow seeded soybean with the dominant *I* allele as compared to pigmented varieties with the recessive *i* allele (Wang et al., 1994). Thus, the biochemical block in pigmentation is via reduction of CHS mRNA and activity. Polymorphisms in CHS genes were found and cosegregated with the *I* locus. Paradoxically, CHS gene duplications suppress and deletions of promoter regions restore expression of the other CHS gene family members (Todd and Vodkin, 1996). These data showed that the *I* locus is a cluster of CHS genes and cosuppression results from naturally occurring duplications leading to lack of pigmentation.

6–6 SUMMARY

Soybean research benefits greatly from a "check-off" program in which the soybean producers have voluntarily chosen to contribute a percentage of their crop proceeds each year to research and marketing. As part of this commitment, the soybean commodity boards have funded an expressed sequence tag (EST) project that will result in approximately 300 000 gene sequence tags being generated and deposited into public databases over the next several years. This will provide the raw material for many gene discovery, evolution, and expression projects including some from the new National Science Foundation Plant Genome Program created in 1998. One of the major objectives of that program is to provide genomic tools for public and private research on economic crop species. The increased rate of discovery from structural and functional genomics research in plants will lead to new products from soybean and production of varieties with improved nutritional and agronomic characteristics through breeding and genetic engineering approaches.

REFERENCES

Akada, S., and S.K. Dube. 1995. Organization of soybean chalcone synthase gene clusters and characterization of a new member of the family. Plant Mol. Biol. 29:189–199.

Akkaya, M.S., A.A. Bhagwat, and P.B. Cregan. 1992. Length polymorphism of simple sequence repeat DNA in soybean. Genetics 132:1131–1139.

Akkaya, M.S., R.C. Shoemaker, J.E. Specht, A.A. Bhagwat, and P.B. Cregan. 1995. Integration of simple sequence repeat DNA markers into soybean linkage map. Crop Sci. 35:1439–1445.

Alderborn, A., A. Kristoffersen, and U. Hammerling. 2000. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res. 10:1249–1258.

Apuya, N., B.L. Frazier, P. Keim., B.J. Roth, and K.G. Lark. 1988. Restriction fragment length polymorphisms as genetic markers in soybean, *Glycine max* (L.) Merr. Theor. Appl. Genet. 75:889–901.

Arabidopsis Genome Initiative 2000. Nature (London) 408:796–815.

Arumuganathan, K., and E.D. Earle. 1991. Nuclear DNA content of some important plant species. Plant Mol. Biol. Rep. 9:208–219.

Bennetzen, J., and M. Freeling. 1993. Grasses as a single genetic system: Genome composition, collinearity and compatibility. Trends Genet. 9:259–261.

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32:314–331.

Bouchez, D., and H. Hofte. 1998. Functional genomics in plants. Plant Physiol. 118:725–732.

Boutin, S.R.Y., N.D. Young, T.C. Olson, Z-H. Yu, R.C. Shoemaker, and C.E. Vallejos. 1995. Genome conservation among three legume genera detected with DNA markers. Genome 38:928–937.

Braun, A., D.P. Little, D. Reuter, B. Muller-Myask, and H. Koster. 1997. Improved analysis of microsatellites using mass spectrometry. Genomics 46:18–23.

Brent, R. 2000. Genomic biology. Cell 100:169–183.

Brookes, A.J. 1999. The essence of SNPs. Gene 234:177–186.

Caetano-Anolles, G., B.J. Bassam, and P.M. Gresshoff. 1992. Primer-template interactions during DNA amplification fingerprinting with single arbitrary oligonucleotides. Mol. Gen. Genet. 235:157–165.

Cahill, D. 2000. High throughput marker assisted selection. p. A02. *In* G.B. Collins et al. (ed.) Proc. 8th Biennial Conf. on the Cellular and Molecular Biology of the Soybean, Lexington, KY. 13–16 Aug. 2000. Univ. of Kentucky, Lexington.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, N. Shaw, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley, and E.S. Lander. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. 22:231–238.

Chen, J., M.A. Iannone, M.S. Li, J.D. Taylor, P. Rivers, A.J. Nelsen, K.A. Slentz-Kesler, A. Roses, and M.P. Weiner. 2000. A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. Genome Res. 10:549–557.

Chen, X., L. Levine, and P.Y. Kwok. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. Genome Res. 9:492–498.

Cho, R.J., M. Mindrinos, D.R. Richards, R.J. Sapolsky, M. Anderson, E. Drenkard, J. Dewdney, T.L. Reuber, M. Stammers, N. Federspiel, A. Theologis, W.H. Yang, E. Hubbell, M. Au, E.Y. Chung, D. Lashkari, B. Lemieux, C. Dean, R.J. Lipshutz, F.M. Ausubel, R.W. Davis, and R.J. Oefner. 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat. Genet. 23:203–207.

Cho, T-J., C.S. Davies, and N.C. Nielsen. 1980. Inheritance and organization of glycinin genes in soybean. Plant Cell 1:329–337.

Clemente, T., B.J. LaValle, A.R. Howe, D.C. Ward, R.J. Rozman, P.B. Hunter, D.L. Broyles, D.S. Kasten, and M.A. Hinchee. 2000. Progeny analysis of glyphosate selected transgenic soybeans derived from Agrobacterium-mediated transformation. Crop Sci. 40:797–803.

Collins, F.S., L.D. Brooks, and A. Chakravarti. 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res. 8:1229–1231.

Cooper, D.N., B.A. Smith, H.J. Cooke, S. Niemann, and J. Schmidtke. 1985. An estimate of unique DNA sequence heterozygosity in the human genome. Hum. Genet. 69:201–205.

Coryell, V.H., H. Jessen, J.M. Schupp, D. Webb, and P. Keim. 1999. Allele-specific hybridization markers for soybean. Theor. Appl. Genet. 98:690–696.

Covitz, P.A., L.S. Smith, and S.R. Long. 1998. Expressed sequence tags from a root-hair-enriched medicago truncatula cDNA library. Plant Physiol. 117(4):1325–1332.

Cregan, P.B., A.A. Bhagwat, M.S. Akkaya, and J. Rongwen. 1994. Microsatellite fingerprinting and mapping of soybean. Methods Cell Mol. Biol. 5:49–61.

Cregan, P.B., and C.V. Quigley. 1997. Simple sequence repeat DNA marker analysis. p. 173–185. *In* G. Caetano-Anolles and P.M. Gresshoff (ed.) DNA markers: Protocols, applications and overviews. John Wiley & Sons, New York.

Cregan P.B., T. Jarvik, A.L. Bush, R.C. Shoemaker, K.G. Lark, A.L. Kahler, N. Kaya, T.T. VanToai, D.G. Lohnes, J. Chung, and J.E. Specht. 1999. An integrated genetic linkage map of the soybean genome. Crop Sci. 39:1464–1490.

Danesh, D., S. Penuela, J. Mudge, R.L. Denny, H. Nordstrom, J.P. Martinez, and N.D. Young. 1998. A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theor. Appl. Genet. 96;196–202.

DellaPenna, D. 1999. Nutritional genomics: Manipulating plant micronutrients to improve human health. Science (Washington DC) 285:375–379.

Delseny, M., R. Cooke, M. Raynal, and F. Grellet. 1997. The arabidopsis thaliana cDNA sequencing projects. FEBS Lett. 405:129–132.

Diers, B.W., P. Keim, W.R. Fehr, and R.C. Shoemaker. 1992a. RFLP analysis of soybean seed protein and oil content. Theor. Appl. Genet. 83;608–612.

Diers, B.W., L. Mansur, J. Imsande, and R.C. Shoemaker. 1992b. Mapping of Phytophthora resistance loci in soybean with restriction fragment length polymorphism markers. Crop Sci. 32:377–383.

DeRisi, J.L., V.R. Iyer, and P.O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science (Washington DC) 278:680–686.

Diwan, N., and P.B. Cregan. 1997. Automated sizing of fluorescent labeled simple sequence repeat markers to assay genetic variation in soybean. Theor. Appl. Genet. 95:723–733.

Ewing, R.M., A.B. Kahou, O. Poirot, F. Lopez, S. Audic, and J.M. Claverie. 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res. 9;950–959.

Ferreira, A.R., K.R. Poutz, and P. Keim. 2000. Soybean genetic map of RAPD markers assigned to an existing scaffold RFLP map. J. Hered. 91:392–396.

Fors, L., K.W. Lieder, S.H. Vavra, and R.W. Kwiatkowski. 2000. Large-scale SNP scoring from unamplified genomic DNA. Pharmacogenomics 1:219–229.

Gilles, P.N., D.J. Wu, C.B. Foster, P.J. Dillon, and S.J. Chanock. 1999. Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips. Nature Biotechnol. 17:365–370.

Gizlice, Z., T.E. Carter, and J.W. Burton. 1994. Genetic base for North American public soybean cultivars released between 1947 and 1988. Crop Sci. 34:1143–1151.

Goldberg, R.B. 1978. DNA sequence organization in the soybean plant. Biochem. Genet. 16:45–68.

Graham, M.A., L.F. Marek, D. Lohnes, P. Cregan, and R.C. Shoemaker. 2000. Expression and genome organization of resistance gene analogs in soybean. Genome 43:86–93.

Grant, D., P. Cregan, and R.C. Shoemaker. 2000. Genome organization in dicots. Genome duplication in Arabidopsis and synteny between soybean and Arabidopsis. Proc. Natl. Acad. Sci. USA 97(8):4168–4173.

Grant, D., and R. Shoemaker. 1997. Molecular hybridization p. 15–26. *In* G. Caetano-Anolles and P.M. Gresshoff (ed.) DNA markers: Protocols, applications and overviews. John Wiley & Sons, New York.

Green, P. 2001. Documentation for Phrap. http://bozeman.mbt.washington.edu, Genome Center, Univ. of Washington, Seattle.

Gurley, W.B., A.G. Hepburn, and J.L. Key. 1979. Sequence organization of the soybean genome. Biochem. Biophys. Acta 561:167–183.

Gut, I.G. 2001. Automation in genotyping of single nucleotide polymorphisms. Hum. Mutat. 17:475–492.

Halushka, M.K., J.B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz, and A. Chakravarti. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. 22:239–247.

Hatey, F., G. Tosser-Klopp, C. Clouscard-Martinato, P. Mulsant, and F. Gasser. 1998. Expressed sequence tags for genes: A review. Genet. Select. Evol. 30:521–541.

Hegenesch, J.B., K.A. Ching, S. Batalov, A.I. Su, J.R. Walker, Y. Zhou, S.A. Kay, R.G. Schultz, and M.P. Cooke. 2001. A comparison of the Celera and Ensemble predicted gene sets reveals little overlap in novel genes. Cell 106;413–415.

Heiter P., and M. Boguski. 1997. Functional genomics: It's all how you read it. Science (Washington DC) 278:601–602.

Hillier, L.D., G. Lennon, M. Becker, M.F. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. Genome Res. 6:807–828.

Hirschhorn, J.N., P. Sklar, K. Lindblad-Toh, Y.M. Lim, M. Ruiz-Gutierrez, S. Bolk, B. Langhorst, S. Schaffner, R. Winchester, and E.S. Lander. 2000. SBE-TAGS: An array-based method for efficient single-nucleotide polymorphism genotyping. Proc. Natl. Acad. Sci. USA 97:12164–12169.

Hoogendoorn, B., M.J. Owen, P.J. Oefner, N. Williams, J. Austin, and M.C. O'Donovan. 1999. Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. Hum. Genet. 104:89–93.

Hymowitz, T., R.J. Singh, and K.P. Kollipara. 1998. The genomes of the Glycine. Plant Breed. Rev. 16:289–317.

Iannone, M.A., J.D. Taylor, J. Chen, M.S. Li, P. Rivers, K.A. Slentz-Kesler, and M.P. Weiner. 2000. Multiplexed single nucleotide polymorphism genotyping by oligonucleotide ligation and flow cytometry. Cytometry 39:131–140.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature (London) 409:860–921.

Jorgenson, R. 1995. Cosuppression, flower color patterns, and metastable gene expression states. Science (Washington DC) 268:686–691.

Kuwabe, A., and N.T. Miyashita. 1999. DNA variation in the basic chitinase locus (*ChiB*) region of the wild plant *Arabidopsis thaliana*. Genetics 153:1445–1453.

Kuwabe A., K. Yamane, and N.T. Miyashita. 2000. DNA polymorphism at the cytosolic phosphoglucose isomerase (*PgiC*) locus of the wild plant *Arabidopsis thaliana*. Genetics 156:1339–1347.

Keim, P., W. Beavis, J. Schupp, and R. Freestone. 1992. Evaluation of soybean RFLP marker diversity in adapted germplasm. Theor. Appl. Genet. 85:205–212.

Keim, P., B.W. Diers, T.C. Olson, and R.C. Shoemaker. 1990. RFLP mapping in soybean: Association between marker loci and variation in quantitative traits. Genetics 126:735–742.

Keim, P., J.M. Schupp, S.E. Travis, K. Clayton, T. Zhu, L. Shi, A. Ferreira, and D.M. Webb. 1997. A high-density soybean genetic map based on AFLP markers. Crop Sci. 37:537–543.

Keim, P., R.C. Shoemaker, and R.G. Palmer. 1989. Restriction fragment length polymorphism diversity in soybean. Theor. Appl. Genet. 77:786–792.

Kokoris, M., K. Dix, K. Moynihan, J. Mathis, B. Erwin, P. Grass, B. Hines, and A. Duesterhoeft. 2000. High-throughput SNP genotyping with the Masscode system. Mol. Diagn. 5:329–340.

Kraft, R., J. Tardiff, K.S. Krauter, and L.A. Leinwand. 1988. Using mini-prep plasmid DNA for sequencing double stranded templates with Sequenase. Biotechniques. 6:544–549.

Kruglyak, L. 1997. The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. 17:21–24.

Kuittinen, H., and M. Aguade. 2000. Nucleotide variation at the *CHALCONE ISOMERASE* locus in *Arabidopsis thaliana*. Genetics 155:863–872.

Kwok, P.Y. 2000. High-throughput genotyping assay approaches. Pharmacogenomics 1:95–100.

Kwok P.Y. and X. Chen. 1998. Detection of single nucleotide variations. p. 125–134. *In* J.K. Setlow (ed.) Genetic engineering. Vol. 20. Plenum Press, New York.

Kwok, P.Y., Q. Deng, H. Zakeri, and D.A. Nickerson. 1996. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. Genomics 31:123–126.

Lackey, J. 1980. Chromosome numbers in the Phaseoleae (Fabaceae Faboideae) and their relation to taxonomy. Am. J. Bot. 67:595–602.

Lander, E.S., and R.A. Weinberg. 2000. Genomics: Journey to the center of biology. Science (Washington, DC) 287:1777–1782.

Lark, K.G., J.M. Weisemann, B.F. Matthews, R. Palmer, K. Chase, and T. Maculma. 1993. A genetic map of soybean (*Glycine max* L.) using an intraspecific cross of two cultivars: 'Minsoy' and 'Noir 1'. Theor. Appl. Genet. 86:901–906.

Lee, J.M., D. Grant, C.E. Vallejos, and R.C. Shoemaker. 2001. Genome organization in dicots. II. Arabidopsis as a 'bridging species' to resolve genome evolution events among legumes. Theor. Appl. Genet. 103:765–773.

Lee, J.M., A. Bush, J.E. Specht, and R.C. Shoemaker. 1999. Mapping of duplicate genes in soybean. Genome 42:829–836.

Lim, C.O., H.Y. Kim, M.O. Kim, S.I. Lee, W.S. Chung, S.H. Park, I. Hwang, and M.J. Cho. 1996. Expressed sequence tags of chinese cabbage flower bud cDNA. Plant Physiol. 111:577–588.

Lin, J.J., J. Ma, and J. Kuo. 1999. Chemiluminescent detection of AFLP markers. Biotechniques 26:344–348.

Lindblad-Toh, K., E. Winchester, M.J. Daly, D.G. Wang, J.N. Hirschhorn, J.P. Laviolette, K. Ardlie, D.E. Reich, E. Robinson, P. Sklar, N. Shah, D. Thomas, J.B. Fan, T. Gingeras, J. Warrington, N. Patil, T.J. Hudson, and E.S. Lander. 2000. Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nature Genet. 24:381–386.

Litt, M., X. Hauge, and V. Sharma. 1993. Shadow bands seen when typing polymorphic dinucleotide repeats: Some causes and cures. Biotechniques 15:280–284.

Litt, M., and J.A. Luty. 1989. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. 44:397–401.

Little, D.P., T.J. Cornish, M.J. O'Donnell, A. Braun, R.J. Cotter, and H. Kuter. 1997. MALDI on a chip: Analysis of arrays of low-femtomole to subfemtomole quantities of synthetic oligonucleotides and DNA diagnostic products dispensed by a piezoelectric pipet. Anal Chem. 69:4540–4546.

Liu, F., B.C. Zhuang, J.S. Zhang, and S.Y. Chen. 2000. Construction and analysis of soybean genetic map. Yi Chuan Xue Bao 27:1018–1026.

Manger, I.D., A. Hehl, S. Parmley, L. D. Sibley, M. Marra, L. Hillier, R. Waterston, and J.C. Boothroyd. 1998. Expressed sequence tag analysis of the bradyzoite stage of Toxoplasma gondii: Identification of developmentally regulated genes. Infect. Immun. 66(4):1632–1637.

Mano, Y., S. Kawasaki, F. Takaiwa, and T. Komatsuda. 2001. Construction of a genetic map of barley (*Hordeum vulgare* L.) cross 'Azumamugi' x 'Kanto Nakate Gold' using a simple and efficient amplified fragment-length polymorphism system. Genome 44:284–292.

Mansur, L.M., J.H. Orf, K. Chase, T. Jarvik, P.B. Cregan, and K.G. Lark. 1996. Genetic mapping of agronomic traits using recombinant inbred lines of soybean [*Glycine max* (L.) Merr.]. Crop Sci. 36:1327–1336.

Marek, L.F., J. Mudge, L. Darnielle, D. Grant, N. Hanson, M. Paz, Y. Huihuang, R. Denny, K. Larson, D. Foster-Hartnett et al. 2001. Soybean genomic survey: BAC-end sequences near RFLP and SSR markers. Genome 44:572–581.

Marek, L.F., and R.C. Shoemaker. 1997. BAC contig development by fingerprint analysis in soybean. Genome 40:420–427.

Marra, M., L. Hillier, T. Kucaba, M. Allen, R. Barstead, C. Beck, A. Blistain, M. Bonaldo, Y. Bowers, L. Bowles et al. 1999. An encyclopedia of mouse genes. Nature Genet. 21(2):1991–1994.

Marshall, A., and J. Hodgson. 1998. DNA chips: An array of possibilities. Nature Biotechnol. 16:27–31.

Martienssen, R., and W.R. McCombie. 2001. The first plant genome. Cell 105:571–574.

Matthews, B.F., T.E. Devine, J.M. Weisemann, H.S. Beard, K.S. Lowers, M.H. MacDonald, Y.B. Park, R. Maiti, J.J. Lin, et al. 2001. Incorporation of sequenced cDNA and genomic markers into the soybean genetic map. Crop Sci. 41:516–521.

Maughan, P.J., M.A. Saghi Maroof, and G.R. Buss. 1995. Microsatellite and amplified sequence length polymorphisms in cultivated and wild soybean. Genome 38:715–723.

Mayer, K., C. Schuller, R. Wambutt, G. Murphy, G. Volckaert, T. Pohl, A. Dusterhoft, W. Stiekema, K. D. Entian, N. Terryn et al. 1999. Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. Nature (London) 402(6763):769–777.

Mazur, B., E. Krebbers, and S. Tingey. 1999. Gene discovery and product development for grain quality traits. Science (Washington DC) 285:372–375.

Meksem, K., B. Ruben, D. Hyten, K. Triwitayakorn, and D A. Lightfoot. 2001. Conversion of AFLP bands into high-throughput DNA markers. Mol. Genet. Genomics 265:207–214

Mian, M.A.R., T. Wang, D.V. Phillips, J. Alvernaz, and H.R. Boerma. 1999. Molecular mapping of the Rcs3 gene for resistance to frogeye leaf spot in soybean. Crop Sci. 39:1687–1691.

Michelmore, R.W., I. Paran, and R.V. Kesseli. 1991. Identification of markers linked to disease resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. USA 88 9828–9832.

Morgante, M., I. Jurman, L. Shi, T. Zhu, P. Keim, and J.A. Rafalski. 1997 The STR120 satellite DNA of soybean: Organization, evolution and chromosomal specificity Chromosome Res. 5:363–373.

Morgante, M., and A.M. Olivieri. 1993. PCR-amplified microsatellites as markers in plant genetics. Plant J. 3:175–182.

Morgante, M., J.A. Rafalski, P. Biddle, S. Tingey, and A.M. Olivieri. 1994. Genetic mapping and variability of seven soybean simple sequence repeat loci. Genome 37:763–769.

Muehlbauer, G.J., J.E. Specht, M.A. Thomas-Compton, P.E. Staswick, and R.L. Bernard. 1988. Near isogenic lines—A potential source in the integration of conventional and molecular marker linkage maps. Crop Sci. 28:729–735.

Muehlbauer, G.J., P.E. Staswick, J.E. Specht, G.L. Graef, R.C. Shoemaker, and P. Keim. 1991. RFLP mapping using near-isogenic lines in the soybean [*Glycine max* (L.) Merr.] Theor. Appl. Genet. 81:189–198.

Mullis, K., F. Faloona, S. Scharf, R. Saiki, G. Horn, and H. Erlich. 1986 Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. Cold Spring Harbor Symp. Quant. Biol 51:263-273.

Narvel, J.M., W.-C. Chu, W.R. Fehr, P.B. Cregan, and R.C. Shoemaker. 2000a. Development of multiplex sets of simple sequence repeat DNA markers covering the soybean genome. Mol. Breed. 6:175–183.

Narvel, J.M., W.R. Fehr, W.-C. Chu, D. Grant, and R.C. Shoemaker. 2000b. Simple sequence repeat diversity among soybean plant introductions and elite genotypes Crop Sci. 40:1452–1458.

Ohno, S. 1970. Evolution by gene duplication. Springer Verlag, New York.

Palmer, R.G., and T.C. Kilen, 1987: Qualitative genetics and cytogenetics. p. 135–209. *In* J.R. Wilcox (ed.) Soybeans: Improvement, production and uses. 2nd ed. ASA, CSSA, and SSSA, Madison, WI.

Palmer, R.G., and R.C. Shoemaker. 1998. Soybean genetics p. 45–82. *In* M. Vidic and D. Jockovic (ed.) Soja. Soybean Inst. of Field and Vegetable Crops, Novi Sad, Yugoslavia.

Prabhu, R.R., and P.M.Gresshoff. 1994. Inheritance of polymorphic markers generated by DNA amplification fingerprinting and their use as genetic markers in soybean. Plant. Mol. Biol. 26;105–116.

Prince, J.A., L. Feuk, W.M. Howell, M. Jobs, T. Emahazion, K. Blennow, and A.J. Brookes. 2001. Robust and accurate single nucleotide polymorphism genotyping by dynamic allele-specific hybridization (DASH): Design criteria and assay validation. Genome Res. 11:152–162.

Purugganan, M.D., and J.I. Suddith. 1999. Molecular population genetics of floral homeotic loci. Departures from the equilibrium-neutral model at the *APETALA3* and *PISTILLATA* genes of *Arabidopsis thaliana*. Genetics 151;839–848.

Putney, S.D., W.C. Herlihy, and P. Schimmel. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. Nature (London) 302(5910):718–721.

Qi, X., S. Bakht, K.M. Devos, M.D. Gale, and A. Osbourn. 2001. L-RCA (ligation-rolling circle amplification): A general method for genotyping of single nucleotide polymorphisms (SNPs). Nucleic Acids Res. 29:E116.

Rafalski, A. 1997. Randomly amplified polymorphic DNA (RAPD) analysis. p. 75–83. *In* G. Caetano-Anolles and P.M. Gresshoff (ed.) DNA markers: Protocols, applications and overviews. J. Wiley & Sons, New York.

Rafalski, A., and S. Tingey. 1993. RFLP map of soybean (*Glycine max*). p. 6.149–6.156. *In* S.J. O'Brien (ed.) Genetic maps: Locus maps of complex genomes. Cold Spring Harbor Lab. Press, New York.

Rector, B.G., A. Demirbas, M.J. Livingston, H.L. Olsen, R.A. Ritchie, G.L. Graef, and J.E. Specht. 1999. Integration of the soybean microsatellite and classical marker maps. P. 136. *In* Proc. Plant and Animal Genome VII. 17–21 Jan. 1999. Scherago Int., New York.

Remington D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize. genome. Proc. Natl. Acad. Sci. USA. 98:11479–11484.

Riley, P.A. 1999. USDA expects record soybean supply. Inform 10(5):503–506.

Rongwen, J., M.S. Akkaya, U. Lavi, and P.B. Cregan. 1995. The use of microsatellite DNA markers for soybean genotype identification. Theor. Appl. Genet. 90:43–48.

Ross, P., L. Hall, I. Smirnov, and L. Haff. 1998. High level multiplex genotyping by MALDI-TOF mass spectrometry. Nature Biotech. 16;1347–1351.

Salanoubat, M., K. Lemcke, M. Rieger, W. Ansorge, M. Unseld, B. Fartmann, G. Valle, H. Blocker, M. Perez-Alonso, B. Obermaier et al. 2000. Sequence and analysis of chromosome 3 of the plant Arabidopsis thaliana. Nature (London) 408(6814):820–822.

Salimath, S., and M.K. Bhattacharyya. 1999. Generation of a soybean BAC library, and identification of DNA sequences tightly linked to the Rps1-k disease resistance gene. Theor. Appl. Genet. 98:712–720.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. Molecular cloning: A laboratory manual. 2nd ed. Cold Spring Harbor Lab., Cold Spring Harbor, New York.

Sapolsky, R., L. Hsie, A. Berno, G. Ghandour, M. Mittman, and J.B. Fan. 1999. High-throughput polymorphism screening and genotyping with high-density oligonucleotide arrays. Genet. Anal. Biomolecular Eng. 14:187–192.

Scallon, B.J., C.D. Dickinson, and N.C. Nielsen. 1987. Characterization of a null-allele for the G_4 glycinin gene from soybean. Mol. Gen. Genet. 208:107–113.

Schena, M., D. Shalon, R.W. Davis, and P.O. Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (Washington DC) 270:467–470.

Schena, M., D. Shalon, R. Heller, A. Chai, P.O. Brown, and R.W. Davis. 1996. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA 93;10614–10619.

Schenk, R.M., K. Kazan, I. Wilson, J.P. Anderson, T. Richmond, S.C. Somerville, J.M. Manners. 2000. Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc. Natl. Acad. Sci. 97;11655–11660.

Shi, M.M. 2001. Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. Clin. Chem. 47:164–172.

Shoemaker, R., P. Keim, L. Vodkin, E. Retzel, S.W. Clifton, R. Waterston, D. Smoller, V. Coryell, A. Khanna, J. Erpelding, X. Gai, V. Brendel, C. Raph-Schmidt, E.G. Shoop, C.J. Vielweber, M.

Schmutz
A compil
Shoemaker, R.C
Shoemaker, R.C
6,131–6.
Spring H
Shoemaker, R.C.
Wilcox, J
(*Glycine*
Shoemaker, R.C.
age group
Singh, R.J., and T
soja Sieb
76:705–7
Somerville, C. 20
Somerville, C.,
285:380–
Song, Q., C.V. Q
selected s
Plant Vari
Sosnowski, R.G.,
gle base
Sci. USA.
Southern, E.M. I
trophoresi
SoyStats: A refere
MO.
Steemers, F.J., J.A
ordered lit
Sterky, F., S. Regu
son, R. Vil
son, M. Ul
sues of pe
95(22):13
Syvanen, A.C. 20
Rev. Genet
Tabata S., T. Kane
Kimura, T.
thaliana. N
Tajima, F. 1983.
105;437–4
Tanabe, K., S. Nak
1999. Expr
ripheral ne
Tenaillon, M.I., M
DNA seque
Acad. Sci.
Theologis, A., J.R.
C.L. Bowm
thaliana. N
Thibaud-Nissen, F.
voals transc
(In press)
Todd, J.J., and Vod
a chalcone
Tomkins, J.P, R. M
ial artificial
with cyst ne
Travis, S.E., J. Mus
phylax var.
5:735–745

Schmutz, D. Pape, Y. Bowers, B. Theising, J. Martin, M. Dante, T Wylie, and C. Granger. 2002. A compilation of soybean ESTs: generation and analysis. Genome 45:329–338.

Shoemaker, R.C. 1999. Soybean genomics from 1985–2002. AgBiotechNet 1:1–4.

Shoemaker, R.C., and T.C. Olson. 1993. Molecular linkage map of soybean (*Glycine max* L. Merr.). p. 6.131–6.138. *In* S.J. O'Brien (ed.) Genetic maps: Locus maps of complex genomes. Cold Spring Harbor Lab. Press, New York.

Shoemaker, R.C., K. Polzin, J. Labate, J. Specht, E.C. Brummer, T. Olson, N. Young, V. Concibido, J. Wilcox, J.P. Tamulonis, G. Kochert, and H.R. Boerma. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–338.

Shoemaker, R.C., and J.E. Specht. 1995. Integration of the soybean molecular and classical genetic linkage groups. Crop Sci 35:436–446.

Singh, R.J., and T. Hymowitz. 1988. The genomic relationship between *Glycine max* (L.) Merr. and *G. soja* Sieb. and Zucc. as revealed by pachytene chromosomal analysis. Theor. Appl. Genet. 76:705–711.

Somerville, C. 2000. The twentieth century trajectory of plant biology. Cell 100:13–25.

Somerville, C., and S. Somerville. 1999. Plant functional genomics. Science (Washington DC) 285:380–383.

Song, Q., C.V. Quigley, T.E. Carter, R.L. Nelson, H.R. Boerma, J. Strachan, and P.B. Cregan. 1999. A selected set of trinucleotide simple sequence repeat markers for soybean variety identification. Plant Varieties Seeds 12:207–220.

Sosnowski, R.G., E. Tu, W.F. Butler, J.P. O'Connell, and M.J. Heller. 1997. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. Proc. Natl. Acad. Sci. USA. 94:1119–1123.

Southern, E.M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98:503–517.

SoyStats: A reference guide to important soybean facts and figures. 1997. Am. Soybean Assoc., St. Louis, MO.

Steemers, F.J., J.A. Ferguson, and D.R. Walt. 2000. Screening unlabeled DNA targets with randomly ordered fiber-optic gene arrays. Nat. Biotechnol. 18:91–94.

Sterky, F., S. Regan, J. Karlsson, M. Hertzberg, A. Rohde, A. Holmberg, B. Amini, R. Bhalerao, M. Larsson, R. Villarroel, M. Van Montagu, G. Sandberg, O. Olsson, T.T. Teeri, W. Boerjan, P. Gustafson, M. Uhlen, B. Sundberg, and J. Lundeberg. 1998. Gene discovery in the wood-forming tissues of poplar: Analysis of 5, 692 expressed sequence tags. Proc. Natl. Acad. Sci. USA 95(22):13330–13335.

Syvanen, A.C. 2001. Accessing genetic variation: Genotyping single nucleotide polymorphisms. Nat. Rev. Genet. 2:930–942.

Tabata S., T. Kaneko, Y. Nakamura, H. Kotani, T Kato, E. Asamizu, N. Miyajima, S. Sasamoto, T Kimura, T. Hosouchi et al. 2000. Sequence and analysis of chromosome 5 of the plant Arabidopsis thaliana. Nature (London) 408(6814):823–826.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tanabe, K., S. Nakagomi, S. Kiryu-Seo, K. Namikawa, Y. Imai, T. Ochi, M. Tohyama, and H. Kiyama. 1999. Expressed-sequence-tag approach to identify differentially expressed genes following peripheral nerve axotomy. Molecular Brain Res. 64:34–40.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.) Proc. Natl. Acad. Sci. USA 98:9161–9166.

Theologis, A., J.R. Ecker, C.J. Palm, N.A. Federspiel, S. Kaul, O. White, J. Alonso, H. Altafi, R. Araujo, C.L. Bowman et al. 2000. Sequence and analysis of chromosome 1 of the plant Arabidopsis thaliana. Nature (London) 408(6814):816–820.

Thibaud-Nissen, F., R.T. Shealy, A. Khanna, and L.O. Vodkin. 2003. Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol. 132. (In press.)

Todd, J.J., and Vodkin, L.O. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. Plant Cell 8:687–699.

Tomkins, J.P., R. Mahalingam, H. Smith, J.L. Goldsbrough, H.T. Knap, and R.A. Wing. 1999. A bacterial artificial chromosome library for soybean PI 437654 and identification of clones associated with cyst nematode resistance. Plant Mol. Biol. 41(1):25–32.

Travis, S.E., J. Maschinski, and P. Keim. 1996. An analysis of genetic variation in *Astragalus cremnophylax* var. *cremnophylax*, a critically endangered plant, using AFLP markers. Mol. Ecol. 5:735–745.

Tyagi, S., D.P. Bratu, and F.R. Kramer. 1998. Multicolor molecular beacons for allele discrimination. Nature Biotech 16:49–53.

Velculescu, V.E., Ahang, L., Vogelstein, B., Kinzler, K.W. 1995. Serial analysis of gene expression. Science (Washington DC) 270:484–487.

Vodkin, L.O., A. Khanna, S. Clough, R. Shenly, R. Philip, J. Erplending, M. Paz, R. Shoemaker, V. Coryell, J. Schupp, P. Keim, A. Rodriquez-Huete, P. Zeng, J. Polnaco, J. Mudge, R. Denny, N. Young, C. Raph, L. Shoop, E. Retzel. 2003. Structural and functional genomics projects in soybean. Plant Mol. Biol. Rep. Supplement 18:2, pS1.

Vos, P., R. Hogers, M. Bleeker, M. Rijans, T. van der Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau. 1995. AFLP: A new technique for DNA fingerprinting. Nucleic Acids Res. 23:4407–4414.

Walbot, V. 1999. Genes, genomes, genomics. What can plant biologists expect from the 1998 National Science Foundation Plant Genome Research Program. Plant Physiol. 119:1151–1155.

Wang, C., Todd, J.J., and Vodkin, L.O. 1994. Chalcone synthase mRNA and activity are reduced in yellow soybean seed coats with dominant *I* alleles. Plant Physiol. 105:739–748

Wang, D.G., J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science (Washington DC) 280:1077–1082.

Watterson, G.A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Pop. Biol. 7:256–276.

Weber, J.L., and P.E. May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. 44:388–396.

Welsh, J., and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. 18:7213–7218.

Williams, J.K.G., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 18:6531–6535.

Wolffe, A.P., and M.A. Matzke. 1999. Epigenetics: Regulation through repression. Science (Washington DC) 286:481–486.

Wu, X.L., C.Y. He, Y.J. Wang, Z.Y. Zhang, Y. Dongfang, J.S. Zhang, S.Y. Chen, and J.Y. Gai. 2001. Construction and analysis of a genetic linkage map of soybean. Yi Chuan Xue Bao 28:1051–1061.

Xing, L., C. Ge, R. Zeltser, G. Maskevitch, B. J. Mayer, and K. Alexandropoulos. 2000. c-Src signaling induced by the adapters Sin and Cas is mediated by Rap1 GTPase. Mol. Cell Biol. 20(19):7363–7377.

Xu, S., R. Singh, and T. Hymowitz. 1998. Establishment of a cytogenetic map of soybean: Current status. Soybean Genet. Newsl. 25:120–122.

Xue, Z.T., M.L. Xu, W. Shan, N.L. Zhuang, W.M. Hu, and S.C. Shen. 1992. Characterization of a Gy_4 glycinin gene from soybean *Glycine max* cv. Forrest. Plant Mol. Biol. 18:897–908.

Yamamoto, E., H.C. Karakaya, and H.T. Knap. 2000. Molecular characterization of two soybean homologs of Arabidopsis thaliana CLAVATA1 from the wild type and fasciation mutant. Biochim. Biophys. Acta 1491:333–340.

Yamamoto, E., and H.T. Knap. 2001. Soybean receptor-like protein kinase genes: Paralogous divergence of a gene family. Mol. Biol. Evol. 18(8):1522–1531.

Yamanaka, N., S. Ninomiya, M. Hoshi, Y. Tsubokura, M. Yano, Y. Nagamura, T. Sasaki, and K. Harada. 2001. An informative linkage map of soybean reveals QTLs for flowering time, leaflet morphology and regions of segregation distortion. DNA Res. 8:61–72.

Young, W.P., J.M. Schupp, and P. Keim. 1999. DNA methylation and AFLP marker distribution in the soybean genome. Theor. Appl. Genet. 99:785–792.

Yu, C.J., Y. Wan, H. Yowanto, J. Li, C. Tao, M.D. James, C.L. Tan, G.F. Blackburn, and T.J. Meade. 2001. Electronic detection of single-base mismatches in DNA with ferrocene-modified probes. J. Am. Chem. Soc. 123:11155–11161.

Yu, Y.G., M.A. Saghai Maroof, G.R. Buss, P.J. Maughan, and S.A. Tolin. 1994. RFLP and microsatellite mapping of a gene for soybean mosaic virus resistance. Phytopathology 84:60–64.

Zakharova, E.S., S.M. Epishin, and Y.P. Vinetski. 1989. An attempt to elucidate the origin of cultivated soybean via comparison of nucleotide sequences encoding glycinin B_4 polypeptide of cultivated soybean, *Glycine max*, and its presumed wild progenitor, *Glycine soja*. Theor. Appl. Genet. 78:852–856.

Zhang, H., L. Yu, N. Mao, Q. Fu, Q. Tu, J. Gao, and S. Zhao. 1999. Cloning, characterization, and chromosome mapping of RPS6KC1, a novel putative member of the ribosome protein S6 kinase family, to chromosome 12q12–q13.1. Genomics 61(3):314–318

Zhu, T., J.M. Schupp, A. Oliphant, and P. Keim. 1994. Hypomethylated sequences: Characterization of the duplicate soybean genome. Mol. Gen. Genet. 244:638–645.

Zhu, T., L. S
DNA
Zhu, T., I. Sl
mole
Zhu, Y.-L., Q
Ficki
bean

EXHIBIT

C

Theor Appl Genet (2004) 109:122–128
DOI 10.1007/s00122-004-1602-3

ORIGINAL PAPER

Q. J. Song · L. F. Marek · R. C. Shoemaker · K. G. Lark · V. C. Concibido · X. Delannay · J. E. Specht · P. B. Cregan

A new integrated genetic linkage map of the soybean

Received: 30 October 2003 / Accepted: 8 January 2004 / Published online: 27 February 2004

Abstract A total of 391 simple sequence repeat (SSR) markers designed from genomic DNA libraries, 24 derived from existing GenBank genes or ESTs, and five derived from bacterial artificial chromosome (BAC) end sequences were developed. In contrast to SSRs derived from EST sequences, those derived from genomic libraries were a superior source of polymorphic markers, given that the mean number of tandem repeats in the former was significantly less than that of the latter ($P<0.01$). The 420 newly developed SSRs were mapped in one or more of five soybean mapping populations: 'Minsoy' × 'Noir 1', 'Minsoy' × 'Archer', 'Archer' × 'Noir 1', 'Clark' × 'Harosoy', and A81-356022 × PI468916. The JoinMap software package was used to combine the five maps into an integrated genetic map spanning 2,523.6 cM of Kosambi map distance across 20 linkage groups that contained 1,849 markers, including 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical traits, six AFLPs, ten isozymes, and 12 others. The number of new SSR markers added to each linkage group ranged from 12 to 29. In the integrated map, the ratio of SSR marker number to linkage group map distance did not differ among 18 of the 20 linkage groups; however, the SSRs were not uniformly spaced over a linkage group, clusters of SSRs with very limited recombination were frequently present. These clusters of SSRs may be indicative of gene-rich regions of soybean, as has been suggested by a number of recent studies, indicating the significant association of genes and SSRs. Development of SSR markers from map-referenced BAC clones was a very effective means of targeting markers to marker-scarce positions in the genome.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s00122-004-1602-3

Communicated by C. Möllers

Q. J. Song · P. B. Cregan (✉)
Soybean Genomics and Improvement Laboratory,
USDA-ARS, Beltsville, MD 20705, USA
e-mail: creganp@ba.ars.usda.gov
Tel.: +1-301-5045070
Fax: +1-301-5045728

L. F. Marek · R. C. Shoemaker
Department of Agronomy, USDA-ARS-CICG,
Iowa State University, Ames, IA 50011, USA

R. C. Shoemaker
Department of Agronomy,
Iowa State University, Ames, IA 50011, USA

K. G. Lark
Department of Biology,
University of Utah,
Salt Lake City, UT 84112, USA

V. C. Concibido · X. Delannay
Monsanto Company,
800 Lindbergh Boulevard, St. Louis, MO 63167, USA

J. E. Specht
Department of Agronomy, University of Nebraska,
Lincoln, NE 68583-0915, USA

Introduction

The first soybean (*Glycine max* L. Merr.) genetic linkage map of molecular markers was reported by Keim et al. (1990). This map consisted of 26 genetic linkage groups containing a total of 150 restriction fragment length polymorphism (RFLP) loci and was based on a F_2 population derived from an interspecific cross of *G. max* (A81-356022) × *G. soja* (PI468916). Lark et al. (1993) subsequently used 132 RFLP, isozyme, and morphological markers to construct a soybean genetic map comprised of 31 linkage groups. Shoemaker and Specht (1995) mapped 110 RFLP, eight random amplified polymorphic DNA (RAPD), seven pigmentation, six morphological, and seven isozyme markers in an F_2 population derived from a mating of isolines of the important soybean cultivars 'Clark' and 'Harosoy'.

These early genetic maps were primarily based on RFLP markers. Due to the lack of polymorphism of RFLP loci in soybean and/or the complexity of multiple DNA banding patterns detected with most RFLP probes, simple sequence repeat (SSR) or microsatellite markers were proposed for map development (Akkaya et al. 1992). Most SSRs are single-locus markers, and many SSR loci are multi-allelic. These characteristics make SSRs an ideal marker system not only for creating genetic maps, but also as an unambiguous means of defining linkage group homology across mapping populations. In 1999, Cregan et al. (1999a) reported the development of 606 SSR loci which, together with 689 RFLP, 79 RAPD, 11 AFLP, ten isozyme, and 26 classical loci, were mapped to one or more of three populations: the USDA/Iowa State *G. max* × *G. soja* F_2, the University of Utah 'Minsoy' × 'Noir 1' recombinant inbred lines, and the University of Nebraska 'Clark' × 'Harosoy' F_2 population. These three separate maps provided useful information relative to the consistency of marker order and genetic distance among the different populations. The Cregan et al. (1999a) report established, for the first time, 20 consensus linkage groups, which were assumed to be the genetic correlates of the 20 soybean chromosomes. In that report, a total of 412 SSR loci were positioned in the 'Minsoy' × 'Noir 1' mapping population of 240 recombinant inbred lines. The resulting map was approximately 2,400 cM in length, but contained 36 intervals of at least 20 cM, and 79 intervals of at least 10 cM, in which no microsatellite loci were positioned. Inversely, there were 67 distinct intervals with less than 0.01 cM of distance between two or more adjacent SSR markers. In some of the 67 intervals, there was no recombination between adjacent SSR loci.

To develop microsatellite markers targeted to SSR-free regions as well as to saturate genomic regions of scientific interest, bacterial artificial chromosome (BAC) libraries can be screened by DNA hybridization or by PCR to identify clones from specific regions of the genome. New SSR or other DNA markers can be subsequently developed from those BAC clones, making it feasible to discover new SSRs associated with RFLP or other previously mapped markers. Employing this strategy, Cregan et al. (1999b) successfully developed new SSR markers targeted to two regions of the soybean genome near soybean cyst nematode-resistance loci on linkage groups G and A2. Genetic mapping confirmed that the new SSRs mapped to the correct sites in the genome.

Genetic markers are frequently polymorphic in one population, but monomorphic in another. JoinMap analysis (Stam 1993; Van Ooijen and Voorrips 2001) allows one to combine data from map populations in which not all markers are in common to obtain combined estimates of recombination. This approach not only increases the number of markers on the map, but also increases map precision and resolution.

In the early stages of microsatellite marker development, genomic DNA fragments containing SSRs were isolated from genomic libraries. More recently, EST sequencing projects have resulted in a wealth of sequence DNA information in numerous crop species including soybean. Some ESTs contain di- and trinucleotide-repeat motifs, making EST collections a potential source of microsatellite markers. The use of ESTs as a source of SSRs has been reported in a number of crop species including rice (Cho et al. 2000), grape (Scott et al. 2000), barley (Kota et al. 2001), sugarcane (Cordeiro et al. 2001), and wheat (Eujayl et al. 2002).

The objectives of the work reported here were: (1) to evaluate the potential of soybean ESTs as a source of SSRs for marker development; (2) to assess the success with which the development of SSR markers could be targeted to specific positions in the soybean genome; (3) to develop an additional set of SSR markers to further saturate the soybean linkage map; and (4) to create a consensus linkage map from five commonly used soybean populations using a JoinMap analysis. The creation of a high-density, integrated soybean linkage map with more precisely positioned markers would permit a better overall assessment of the distribution of SSR loci in the soybean genome. Moreover, the map would be useful for map-based cloning efforts and would provide a framework for the positioning of single nucleotide polymorphism (SNP)-based loci that are currently being developed from existing ESTs and other available sources of DNA sequence (Zhu et al. 2003).

Materials and methods

Sources of SSR-containing sequences

Random genomic DNA

The basic procedures of cloning and identification of microsatellite-containing, 500–700-bp genomic clones of 'Williams' soybean DNA were described previously (Cregan et al. 1994; Akkaya et al. 1995). Primer pairs were designed for the flanking regions of repeat motifs that consisted of either ten or more dinucleotide repeat units, or eight or more trinucleotide repeat units.

Targeted SSR-marker development

BAC clones putatively associated with specific positions in the soybean genome were identified either by hybridization of RFLP probes (Marek and Shoemaker 1997) or via PCR as suggested by Green and Olsen (1990). RFLP probes were used in an attempt to identify BAC clones at genome locations where RFLP loci, but no SSR loci, were present. Conversely, SSRs were used to identify BAC clones in an attempt to develop additional SSR markers targeted to a specific genomic location. The details relating to the use of BAC clones as a source of DNA for targeted-SSR development were described by Cregan et al. (1999b).

SSR-containing repeats from ESTs

Upon the initiation of this project (December 2000), 136,800 soybean ESTs were available in GenBank. These ESTs were screened to identify sequences containing ten or more dinucleotide SSRs or eight or more trinucleotide SSRs.

Primer design and examination

PCR primers were designed to the flanking regions of microsatellites with ten or more dinucleotide and eight or more trinucleotide repeats using the software program Oligo 5.0 (National Biosciences, Plymouth, Minn.). Primers were synthesized by BioServe Biotechnologies (Laurel, Md.). Each primer pair was empirically tested for polymorphism using 'Clark', 'Harosoy', 'Jackson', 'Williams', 'Amsoy', 'Archer', 'Fiskeby', 'Minsoy', 'Noir 1', 'Tokyo', A81-356022 (*G. max*) and PI468916 (*G. Soja*) genomic DNA as templates. The first 10 of the above 12 genotypes were described by Cregan et al. (1999a). These ten genotypes represented a range of diversity within the cultivated soybean species. Primers designed from the ESTs were only tested on 'Minsoy', 'Noir 1', and 'Archer'. The ^{32}P-labelled PCR products were analyzed on a 6% DNA sequencing gel with 30% formamide, followed by autoradiography.

Mapping populations

Five widely used soybean mapping populations were used for microsatellite positioning; three of these, the USDA/Iowa State University A81-356022 × PI468916 (MS) population, the University of Nebraska, 'Clark' isoline × 'Harosoy' isoline (CH) population, and University of Utah 'Minsoy' × 'Noir 1' (MN) population, were previously described by Cregan et al. (1999a). The University of Utah 'Minsoy' × 'Archer' (MA) and 'Archer' × 'Noir 1' (AN) RIL populations were described by Mansur et al. (1995, 1996). Newly developed SSRs were mapped to MA, MN, and/or NA populations, and then JoinMap analysis was used on the five populations.

DNA, isozyme, and classical genetic markers

A data set containing 1,019 SSR, 749 RFLP, 13 AFLP, 90 RAPD, ten isozyme, 24 classical, and 12 other markers that mapped in at least one of the five populations CH, MS, MA, MN, and/or AN was used for map integration.

Statistical analysis

Linkage map construction using JoinMap analysis

Linkage maps of the five mapping populations were integrated based on the principle described by Stam (1993) using the JoinMap 3.0 (Van Ooijen and Voorrips 2001) program. The initial step involved calculating the LOD scores and pairwise recombination frequencies between markers. A LOD of 5.0 was used to create linkage groups in the MS, MA, and AN populations, whereas a LOD 4.0 was used in the MN and CH populations. The five maps of each linkage group were then integrated. Recombination values were converted to genetic distances using the Kosambi mapping function. The resulting 20 linkage groups were identified using the alphanumeric codes described in Cregan et al. (1999a).

SSR marker distribution

The theoretical distribution of map distance between adjacent SSR markers was estimated based on the assumption of random distribution of markers over the total length of the linkage map. The goodness of fit between the observed and theoretical distribution was tested using the Monte Carlo estimate of chi-square in Proc-StatXact 5 of SAS (Mehta and Patel 2002). The Monte Carlo estimate of the exact *P*-value was based on a Monte Carlo sample of size 10,000. To avoid the bias, markers developed from targeted isolation of BACs were excluded from this analysis.

Results

Development of SSR markers from EST and genomic DNA sequences

Dinucleotide and trinucleotide SSRs were identified in EST and BAC end sequences from GenBank, BAC subclones, and from clones of genomic libraries. The minimum length criteria were ten or more repeat units for dinucleotide repeats and eight or more for trinucleotide repeats. A total of 420 new SSR loci were developed to add to the 606 SSR loci published by Cregan et al. (1999a). Among these 420 SSRs, 24 were developed from EST sequences, five from GenBank BAC end sequences, 127 from DNA of BAC subclone libraries intended to target specific map positions, and 264 from genomic libraries. Primer pairs designed for sequences with an ATT/TAA, AT/TA, CT/GA, and various other repeat motifs, numbered 110, 276, 12, and 22, respectively.

Of the 136,800 soybean EST sequences examined, 75 contained dinucleotide repeats of ten or more, and 58 ESTs contained trinucleotide repeats of eight or more. The average percentage of ESTs containing the minimum number of repeats was thus less than 0.1%. Of the 133 primer sets designed for the EST-derived SSRs, just 24 (18.0%) amplified polymorphic products among the genotypes of 'Minsoy', 'Noir 1', and 'Archer' (Table 1). In contrast, over the course of several years of SSR-marker development in soybean, 824 (43%) primer sets designed for SSRs derived from genomic libraries were polymorphic among these three genotypes. This proportion was significantly higher than the observed polymorphism rate from the EST-derived primer sets (t=6.34, P<0.01). The mean length of di- and trinucleotide repeats was also significantly shorter (t=5.7, P<0.01 and t=9.3, P<0.01 for di- and trinucleotide repeats, respectively) in the EST-derived SSRs compared to the SSRs from genomic DNA sequences (Table 1).

Table 1 Means and standard deviations (SD) of repeat numbers in simple sequence repeats (*SSRs*) obtained from either ESTs or from genomic DNA sequences

Motif	Primers designed to SSRs from EST sequences			Primers designed to SSRs from genomic DNA sequences		
	No. of primer pairs	Mean(±SD) repeat length	No. of polymorphic loci	No. of primer pairs	Mean(±SD) repeat length	No. of polymorphic loci
Dinucleotide	75	18±6.7	14 (19%)	693	24±7.4	283 (40%)
Trinucleotide	58	11±2.8	10 (17%)	1,211	16±5.9	541 (45%)

Table 2 Number of markers mapped to each linkage group and linkage group length in Kosambi mapping distance

Linkage group	No. SSR		No. RFLP		No. RAPD	No. AFLP	Other	Total	Length (cM)
	Previously mapped	New	Previously mapped	New					
A1	27	23	36	1	-	-	-	87	102.3
A2	37	27	44	2	2	-	4	116	165.7
B1	19	16	32	1	2	-	1	71	131.8
B2	24	12	38	4	6	2	2	88	120.9
C1	21	22	19	4	-		4	70	135.6
C2	35	18	41	3	2	-	1	100	157.9
D1a	39	14	33	4	5	-	6	101	121.0
D1b	30	29	18	1	1	1	2	82	138.0
D2	39	21	18	1	4	1	3	87	133.9
E	28	15	42	5	11	-	2	103	71.3
F	40	24	37	4	4	1	3	113	151.0
G	36	27	50	3	12	-	1	129	116.8
H	21	17	34	7	3	-	2	84	124.0
I	21	19	30	2	2	-	2	76	125.2
J	22	28	31	12	5	-	-	98	91.0
K	40	19	22	2	4	1	4	92	117.0
L	31	21	41	2	2	-	2	99	115.1
M	25	26	22	2	2	-	1	78	142.2
N	24	21	25	4	4	-	4	82	116.7
O	36	21	30	2	2	-	2	93	146.4
Total	595	420	643	66	73	6	46	1849	2,523.6

Targeted SSR marker development

Of the 127 SSR markers developed from BAC subclone libraries, 91 originated from BAC clones identified by existing RFLP probes and 36 from BAC clones identified via existing SSR markers. However, only 36 of the 91 (39.6%) compared to 23 of 36 (64%) markers subsequently mapped to the genomic regions to which they were targeted.

Mapping of the SSR markers

A JoinMap analysis of the 1,019 SSR, 749 RFLP, 13 AFLP, 90 RAPD, ten isozyme, and 30 other markers that segregated in at least one of the five populations produced a genetic map comprised of 20 consensus linkage groups that spanned 2,523.6 cM of Kosambi map distance. A total of 1,849 markers, including 1,015 SSRs, 709 RFLPs, 73 RAPDs, 24 classical traits, six AFLPs, ten isozymes, and 12 others, were integrated to form the current map (Table 2). Four SSRs remained unlinked, as did 40 of the RFLP loci. Among the 1,849 markers, a total of 420 SSR and 66 RFLP have been added to the map since the report by Cregan et al. (1999a). The numbers of SSRs mapped per linkage group averaged about 51, but varied from 35 to 64. The average length of the interval between any two adjacent SSR markers was 2.5 cM. The primer sequences for all SSR loci, as well as genetic maps of each of the 20 consensus linkage groups, are available on the SoyBase Web site of the USDA, ARS Soybean Genome Database (http://soybase.agron.iastate.edu/). Additional details can be found on the corresponding author's Web site http://bldg6.arsusda.gov/~pooley/soy/cregan/soy_map1.html.

Distribution of SSR markers among and within linkage maps

The MN map presented in Cregan et al. (1999a) had 36 intervals of greater than 20 cM in which there was no SSR locus. With JoinMap integration and SSR markers from the other four populations, the gaps in the MN map were filled with 76 markers previously mapped by Cregan et al. (1999a) in the MS and CH populations. In the current study, 90 of the 420 new SSR loci developed either randomly or by targeting mapped to 30 of the 36 intervals. Six of the 36 intervals, C2 Satt202-Satt371; D1a Satt531-Satt368; D1b Satt542-Satt412; H Satt353-Satt192; I top-Satt571; and O Sat_109-Scaa001 still contain no new SSR markers. The number of markers mapped to the remaining 30 intervals varied from one to eight (Table 3).

Chi-square tests of the number of markers mapped to each linkage group indicated a significant deviation from that anticipated based upon linkage group length (χ^2=36.7, P<0.05). However, this deviation was mainly due to fewer and greater numbers of new SSRs mapping to linkage groups B1 and G (Fig. 1). Indeed, a recalculation of the chi-square, with the G and B1 linkage groups excluded from the analysis, indicated similar SSR marker density among the 18 remaining linkage groups.

The randomness of SSR-marker distribution within linkage groups was also examined. Observed and theoretical distributions of map distances between adjacent SSR markers were not completely congruent; the Monte Carlo estimate of the exact P-value based on a Monte Carlo sample of size 10,000 is less than 0.01. As indicated in Fig. 2, there were large differences in the observed and expected frequencies of the cases in which adjacent SSR markers were separated by 0.5 cM or by 1.0 cM. The observed and the expected numbers were

Table 3 Number of new SSR loci mapped to genomic intervals of at least 20 cM that contained no SSR markers in the soybean genetic map reported by Cregan et al. (1999a)

Linkage group	Flanking SSR loci or linkage-group end	No. of previously existing markers positioned via JoinMap analysis	No. of new markers
A1	Satt050-Satt385	0	4
	Satt424-Sat_115	2	1
B1	Top-Satt509	1	4
	Satt197-Satt298	2	2
	Sat_123-Satt453	2	1
B2	Satt577-Satt126	2	2
	Satt126-Sct_034	0	2
	Satt534-Satt560	2	1
C1	Soygpatr-Satt578	1	4
	Sat_042-Satt524	1	6
C2	Sat_130-Sat_062	4	3
	Satt291-Satt170	4	2
	Satt202-Satt371	1	0
D1a	Satt531-Satt368	0	0
	Sat_036-Satt071	2	2
D1b	Sat_096-Satt095	1	3
	Satt542-Satt412	3	0
	Sat_069-Satt459	2	1
D2	Satt301-Sat_086	5	2
E	Satt384-Satt598	9	7
F	Satt522-Sat_074	1	3
G	Satt288-Satt472	0	3
H	Satt353-Satt192	3	0
I	Top-Satt571	0	0
J	Sct046-Satt456	6	8
	Satt215-Satt244	3	3
K	Sat_043-Satt475	1	1
	Satt260-Sat_020	0	4
L	Satt462-Satt481	0	4
M	Satt150-Satt567	0	1
N	Top-Satt159	1	3
	Satt387-Satt521	1	1
O	Satt445-Satt259	0	5
	Satt347-Satt262	10	5
	Satt123-Satt243	6	2
	Sat_109-Scaa001	0	0

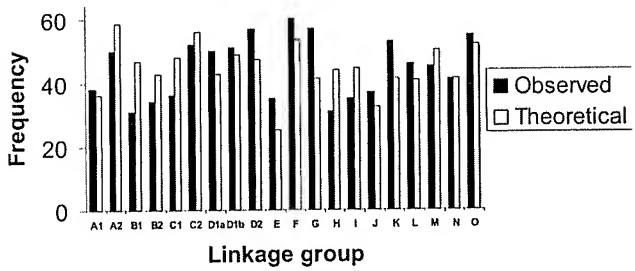

Fig. 1 Observed and theoretical distribution of simple sequence repeat (SSR) markers in linkage groups based on the ratio of mapped SSR markers to linkage group length (cM)

187 versus 133 and 148 versus 113, respectively. These data indicated that more SSRs than average are closely linked, thus suggesting some degree of SSR-marker clustering.

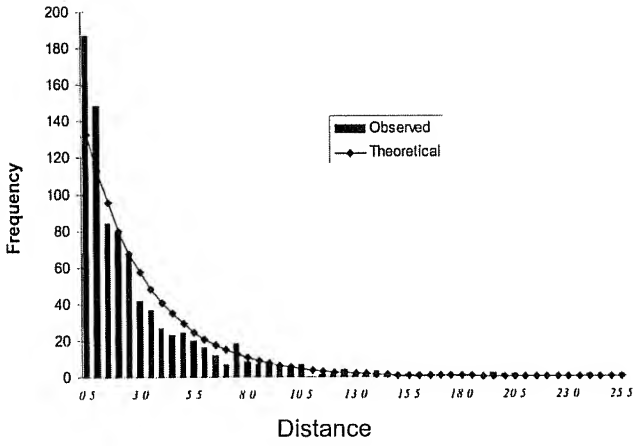

Fig. 2 Theoretical and observed distribution of Kosambi map distance between adjacent SSR markers (summarized over all linkage groups)

Discussion

We designed primers to 133 sequences with microsatellite repeats derived from ESTs, but only 24 (18.0%) of those primer sets produced useful polymorphic markers. In contrast, when genomic DNA sequences were used as the source of SSR-containing sequences, 43.0% yielded markers that were polymorphic with respect to the genotypes of 'Minsoy', 'Noir 1', and 'Archer'. Markers derived from genomic libraries also contained more repeat units as well as a greater range of allele sizes and genetic diversity than markers isolated from EST libraries. The striking difference of polymorphism between the soybean SSRs derived from the two sources is consistent with differences reported in rice (Temnykh et al. 1999; Cho et al. 2000), sugarcane (Cordeiro et al. 2001), tomato (Arshchenkova and Ganal 2002), wheat (Eujayl et al. 2002), and barley (Thiel et al. 2003). For example, Arshchenkova and Ganal (2002) reported that only 20 of 27,000 tomato ESTs contained microsatellites of more than ten repeat units. EST-derived microsatellites were generally shorter (7.3 repeat units) than genomic DNA-derived microsatellites (22.7 repeat units) in barley (Ramsay et al. 2000). The average number of repeats from EST-derived and genomic DNA-derived SSRs was 6.1 versus 13.7 in sugarcane (Cordeiro et al. 2001). The expansion or contraction of dinucleotide repeat length in exons may likely be suppressed due to the deterious nature of the frame-shift mutation that would frequently result in translated regions. Microsatellite markers derived from repeat arrays in genes are reported to be significantly less polymorphic than markers generated from longer arrays (Smulders et al. 1997). Other factors such as selection against large alteration in coding DNA or even a closely associated sequence that may play a role in gene expression could constrain microsatellite expansion or contraction. Such constraints could contribute to the reduced polymorphism of microsatellites in ESTs. To

Table 4 Position of repeat motif in the sequenced genes from which polymorphic SSR markers were developed

GenBank accession number	Description	Motif	Repeat position
AB002807	*Glycine max* DNA for modulin 35	(AT)14	Boundary 5′ upstream sequences
AF162283	*Glycine max* acetyl-CoA carboxylase (*accB-1*) gene	(CT)11	5′ Untranslated region
AF186183	*Glycine max* retrovirus-like element Calypso2-1	(ATT)22	Boundary 5′ upstream sequences
X53404	*Glycine max* glycin A (1a)B(1b) and A(2)B(1a) boundary DNA	(AT)25	Intragenic
X17120	Soybean actin *SAc7* gene	(CT)16	Boundary 5′ upstream sequences
X16876	Soybean *ENOD2B* gene	(AT)17	Intragenic
X01425	Soybean pseudogene for leghemoglobin	A18	Intron
X07159	Soybean pseudogene for heat shock protein Gmshp17.9-D (classVI)	(AT)9	Boundary 5′ upstream sequences
V00458	*Glycine max* gene encoding ribulose-1,5-bisphosphate carboxylase small subunit	A20	Intron
X56139	Soybean *ac514* gene for lipoxygenase	(AT)13	Boundary 5′ upstream sequences
L23833	Soybean glutamine phosphoribosylpyrophosphate amidotransferase	(CTT)6(CTT)4	5′ Untranslated region
M11317	Soybean (*Glycine max*) low MVV heat shock protein gene (Gmhsp17.6-L)	(AT)15	Boundary 5′ upstream sequences
V00452	*Glycine max* leghemoglobin gene	(AT)26	Intron
M94764	*Glycine max* nodulin gene	(AT)24	Intron
J02746	*Glycine max SbPRP1* gene encoding a proline-rich protein	(ATT)20	Boundary 5′ upstream sequences

gain further understanding of the position of SSRs in and around functioning genes, the position of SSRs in the 15 genes from which we have developed polymorphic markers was determined (Table 4). In two instances, the SSR was located in 5′ UTR sequence, while in all others, they were located in either 5′ boundary sequence (seven cases), introns (four cases), or in intragenic sequence (two cases). This suggested that even when polymorphic SSRs were discovered in genic or perigenic regions, the SSR-repeat sequence changes occur only infrequently in mRNA. Obviously, EST-sequence data provide a convenient source of SSR-containing sequences that may be easily and inexpensively exploited. However, even in species with large EST collections, relatively few informative SSR loci are likely to result from this source.

Clustering of SSR markers on the soybean map was observed. Similar clustering of SSR markers was also reported in the tomato (Broun and Tanksley 1996; Areshchenkova and Ganal 1999) and rice linkage maps (McCouch et al. 2003). Physical clustering of SSR markers was also reported in the rat radiation hybrid map (Watanabe et al. 1999) and in barley (Cardle et al. 2000). Morgante et al. (2002) and Cardle et al. (2000) indicated that microsatellites are significantly associated with the low-copy fraction of plant genomes based on the estimation of microsatellite density in *Arabidopsis thaliana*, rice, soybean, maize, and wheat. Among these species, the overall frequency of microsatellites was inversely related to genome size and to the proportion of repetitive DNA. This suggests that most microsatellites reside in regions predating the recent genome expansions in many plants. In order to investigate the distribution of SSRs per megabase (Mb) on each of the 12 rice chromosomes, McCouch et al. (2003) divided the total number of SSRs mapped to each chromosome by the total length of genomic sequence available for each. The figures were compared to the number of EST clusters/Mb on each chromosome identified by The Institute for Genomic Research's *Oryza* gene index using the same genomic sequence data. The density of genes was approximately ten times the density of newly developed SSR markers, but there was a significant correlation (r=0.45, P<0.015) between the number of genes/Mb and the number SSRs/Mb at the level of the chromosome. The clustering of SSR loci we have observed in this report may correspond to gene-rich regions of soybean. However, as a result of the shorter length criteria used to define a microsatellite in studies such as Morgante et al. (2002) versus that used here, this conclusion remains tentative.

Thirty of the 36 intervals in the MN population described by Cregan et al. (1999a) that then contained no SSR marker now have at least one, and often several, SSR markers based on the results of our present study. In many instances, new markers were positioned in these intervals as a result of SSRs obtained from genomic clones, but 127 were developed from BAC clones with the intention of targeting specific genomic intervals. The proportion of SSRs that mapped to the linkage groups to which they were targeted was much higher (64%) when BAC clones were identified using PCR primers to SSR-flanking regions than when the BAC clones were identified with RFLP probes (39.6%). The higher efficiency of targeting when using SSR rather than RFLP probes is likely due to the greater specificity of PCR versus hybridization. The difference is also consistent with the report that RFLP probes hybridize, on average, to 2.55 map positions in the soybean genome (Shoemaker et al. 1996), and that the multiple fragments detected by RFLP frequently occur on different linkage groups (Keim et al. 1990). If only 1 of every 2.55 hybridizations per probe were to the targeted position in the genome, then approximately 39% of the BACs identified would be from the desired position in the genome. Thus, our results suggest that hybridization was about as successful as would be anticipated for the identification of a BAC clone from a specific position in the soybean genome. Although targeting was more

successful when SSRs were used to identify BAC clones, the duplicated nature of the genome still interfered with the efficiency of BAC clone identification despite the greater specificity of PCR. This is likely a reflection of the fact that at least some regions of the soybean genome share very high levels of sequence homology (Zhu et al. 1994; Shoemaker et al. 1996). This may make the development of locus-specific markers to these duplicated regions extremely difficult.

Acknowledgements The authors wish to thank Edward Fickus for excellent technical assistance. The financial support of the United Soybean Board (grant nos. 1243 and 2212) and the Monsanto Company is gratefully acknowledged.

References

Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. Genetics 132:1131–1139

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat (SSR) DNA markers into a soybean linkage map. Crop Sci 35:1439–1445.

Areshchenkova, T, Ganal MW (1999) Long tomato microsatellites are predominantly associated with centromeric regions. Genome 42:536–544

Areshchenkova T, Ganal MW (2002) Comparative analysis of polymorphism and chromosomal location of tomato microsatellite markers isolated from different sources. Theor Appl Genet 104:229–235

Broun P, Tanksley SD (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet 250:39–49

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh (2000) Characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Park WD, Ayres N, Cartinhour S (2000) Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza Sativa* L.). Theor Appl Genet 100:713–722

Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ (2001) Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 160:1115–1123

Cregan PB, Bhagwat AA, Akkaya MS, Jiang R (1994) Microsatellite fingerprinting and mapping of soybean. Methods Mol Cell Biol 5:49–61

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, VanToai TT, Lohnes DG, Chung J, Specht JE (1999a) An integrated genetic linkage map of the soybean. Crop Sci 39:1464–1490

Cregan PB, Mudge J, Fickus ED, Marek LF, Danesh D, Denny R, Shoemaker RC, Matthews BF, Jarvik T, Young ND (1999b) Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor Appl Genet 98:919–928

Eujayl I, Sorrells ME, Baum M, Wolters P (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. Theor Appl Genet 104:399–407

Green ED, Olson MV (1990) Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. Proc Nat Acad Sci USA 87:1213–1217

Lark KG, Weisemann JM, Matthews BF, Palmer R, Chase K, Macalma T (1993) A genetic map of soybean (Glycine max L.) using an intraspecific cross of two cultivars: 'Minsoy' and 'Noir 1'. Theor Appl Genet 86:901–906

Kota R, Varshney RK, Thiel T, Dehmer KJ, Graner A (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). Hereditas 135:145–151

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Mansur LM, Orf J (1995) Evaluation of soybean recombinant inbreds for agronomic performance in northern USA and Chile. Crop Sci 35:422–425

Mansur LM, Orf JH, Chase K, Jarvik T, Cregan PB, Lark KG (1996) Genetic mapping of agronomic traits using recombinant inbred lines of soybean. Crop Sci 36:1327–1336

Marek LF, Shoemaker RC (1997) BAC contig development by fingerprint analysis in soybean. Genome 40:420–427

McCouch SR, Teytelman L, Xu Y, Lobos KB, Clare K, Walton M, Fu B, Maghirang R, Li Z, Xing Y, Zhang Q, Kono I, Yano M, Fjellstrom R, DeClerck G, Schneider D, Cartinhour S, Ware D, Stein L (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). DNA Res 9:257–279

Mehta C, Patel N (1997) Proc-StatXact for SAS users. Cytel, Cambridge, Mass.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Ooijen JW van , Voorrips RE (2001) JoinMap 3.0 software for the calculation of genetic linkage maps. Plant Research International, Wageningen, The Netherlands

Ramsay L, Macaulay M, Ivanissivich S, MacLean K, Cardle L, Fuller J, Edwards K, Tuvesson S, Morgante M, Massari A, Maestri E, Marmiorlin N, Sjakste T, Ganal M, Powell W, Powell W, Waugh R (2000) A simple sequence repeat-based linkage map of barley. Genetics 156:1997–2005

Scott KD, Eggler P, Seaton G, Rosetto M, Ablett EM, Lee LS, Henry RJ (2000) Analysis of SSRs derived from grape ESTs. Theor Appl Genet 100:723–726

Shoemaker RC, Specht JE (1995) Integration of the soybean molecular and classical genetic linkage groups. Crop Sci 35:436–446

Shoemaker RC, Polzin K, Labate J, Specht J, Brummer EC, Olson T, Young N, Concibido V, Wilcox J, Tamulonis JP, Kochert G, Boerma HR (1996) Genome duplication in soybean (*Glycine* subgenus *soja*). Genetics 144:329–339

Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor Appl Genet 94:264–272

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Temnykh S, Park W, Ayres N, Cartinhour S, Hauck N, Lipovich L, Cho YG, Ishii T, McCouch SR (1999) Mapping and genome organization of microsatellites in rice (*Oryza Sativa* L.). Theor Appl Genet 100:698–712

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet 106:411–422

Watanabe TK, Bihoreau MT, McCarthy LC, Kiguwa SL, Hishigaki H, Tsuji A, Browne J, Yamasaki Y, Mizoguchi-Miyakita A, Oga K, Ono T, Okuno S, Kanemoto N, Takahashi E, Tomita K, Hayashi H, Adachi M, Webber C, Davis M, Kiel S, Knights C, Smith A, Critcher R, Miller J, James MR, et al (1999). A radiation hybrid map of the rat genome containing 5,255 markers. Nat Genet 22:27–36

Zhu T, Schupp JM, Oliphant A, Keim P (1994) Hypomethylated sequences: characterization of the duplicate soybean genome. Mol Gen Genet 244:638–645

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134